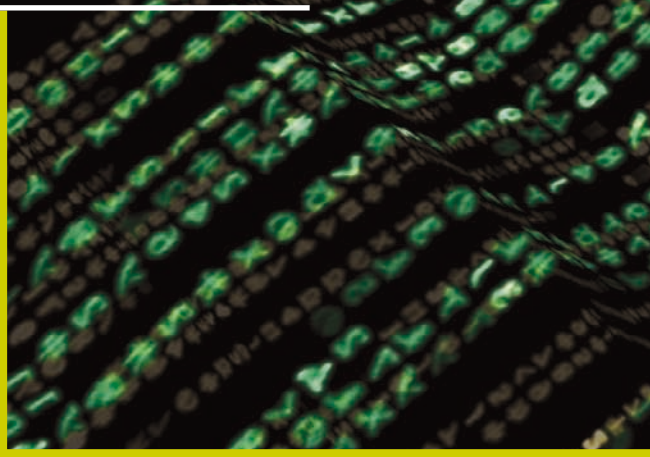
Dealing with the data flood

MINING DATA, TEXT AND MULTIMEDIA

EDITED BY JEROEN MEIJ



STT 65

STT Netherlands
Study Centre for
Technology Trends

Dealing with the data flood

STT Netherlands
Study Centre for
Technology Trends

The Netherlands Study Centre for Technology Trends (STT) was founded in 1968 by the Royal Institution of Engineers in the Netherlands (KIVI). STT/Beweton has the following aims:

- To evaluate technological trends from the viewpoint of the engineering sciences and to explore their interaction with other developments in society as a whole.
- To give wide publicity to its findings as a contribution to a more integrated picture of the future of society in the Netherlands and elsewhere.

STT addresses itself to industry, government, science, and the interested layman.

STT/Beweton is established at Prinsessegracht 23, The Hague.
Correspondence address:
P.O. Box 30424, 2500 GK The Hague, The Netherlands.
Telephone +31 70 302 98 30.
E-mail info@stt.nl

Dealing with the data flood

Mining data, text and multimedia

Edited by Jeroen Meij

2002

STT/Beweton
The Hague, The Netherlands

Colofon

American proofreading Geoff Salvidant M.A.
Editor Jeroen Meij, jeroen.meij@xs4all.nl
Book design Salabim Design Consultancy BNO, Rotterdam
Printing Drukkerij Liesbosch, Nieuwegein

CIP-data Koninklijke Bibliotheek, The Hague

ISBN 90-804496-6-0
NUGI 841
Key words data mining, text mining, multimedia mining, web mining, applications, techniques, overview

Acknowledgements

We owe special thanks to our project sponsors VNU Publishers in Haarlem and Perot Systems Netherlands BV in Amersfoort.

The general STT sponsors are mentioned at the back of this book.

Contents

Prologue

Many companies are already using data mining techniques to approach their target group of potential customers. The same techniques will support the design of the optimal ripening and storage strategy for fruits.
Some of these techniques will also provide a bird's eye view of document collections, including relations between documents. In life sciences data mining will assist to assign functions to genes, and in linking chemical structures (drugs) to biological effects.
Envisat, a new environmental satellite has just been put into orbit. Data mining will help the interpretation and understanding of the high resolution data this satellite will provide.
Specifically this understanding may be the most important effect of the use of data mining tools, an understanding which will enable us to learn and to increase our knowledge base .

The exact value of knowledge has been the topic of much debate the last decades. Early in the 21st century, both the economic and societal value of knowledge are widely recognized. This is not only true in the academic fields, but also for the research environment in industry, or any other 'learning' environment.

This comes as no surprise: by definition, obtaining knowledge requires personal effort, making knowledge an enduring scarcity and therefore a valuable asset.

This book describes new tools and directions that will help us convert something cheap, which is abundantly available — data — into something scarce and valuable: knowledge.

The Hague, April 2002

Chairman STT/Beweton Ir R.M.J. van der Meer

1

Introduction

1.1 Introduction

1.1.1 About this book

This book is the result of a project started in 1999, in which over 80 experts from several areas contributed to create an overview and to generate visions of the future. It will try to address ways of coping with the rapidly growing quantities of data. Data available to us, but also data available about us.
Can sales slips of the past years help us forecast the sales of beer on a rainy day with and without a big sports event? Once I have my holiday movies digitized, can I search them for the frame in which my daughter is wearing this red sweater? Can we use a year of satellite data to generate hypotheses?
We will first sketch a background for the field and identify some relevant trends.

1.1.2 Preface

David J. Hand[1]

With their use of the telescope, Galileo Galilei and others opened the doors to the macroscopic universe. They enabled us to see objects which were so far away that they were invisible to the unaided eye. With their use of the microscope, Antoni van Leeuwenhoek and others opened the doors to the microscopic universe. They enabled us to see objects, which were so small that they were invisible to the naked eye. These instruments, the telescope and microscope, amplified natural human abilities many millionfold, permitting humanity to study and understand things the existence of which we previously could never have even dreamed. This book describes, and illustrates with real case studies, another set of instruments, which enable us to see things we could never perceive with the unaided eye and brain. Telescopes explore gigantic objects, and microscopes explore minuscule objects. The instruments described in this book explore aggregate objects. Aggregate objects are collections of data describing many individual objects. The constituent objects have properties, and one can study such objects singly, but the unassisted mind cannot study an aggregate object as a whole. This would not matter, if the properties possessed by the aggregate object were the same as those possessed by the individual objects. But they are not. Aggregate objects have other properties, often quite different from those of their constituents. And aggregate objects often have properties which their constituents cannot possess.

What sort of things are aggregate objects? A human population is an aggregate object. The collection of purchases by shoppers in a supermarket is an aggregate object. The set of descriptions of all the visible stars is an aggregate object. Descriptions of segments of the human genome form an aggregate object. A collection of paths taken, when surfing the web is an aggregate object. A company's database of credit card transaction records is an aggregate object. A library of extracts from a newspaper is an aggregate object.

And what sort of properties do aggregate objects possess? In particular, what sort of properties do such objects possess that their constituents can not? A human population can have several different kinds of individuals within it, but a single individual is of only one type. The collection of purchases by shoppers in a supermarket may enable one to predict how new customers will behave, but such a prediction cannot be made merely by observing one shopper. By studying a collection of stars, we can develop a theory about the natural life stages of a star, but, short of watching for billions of years, this cannot be done by observing a single star. By studying similarities and differences between genome

1 Prof D.J. Hand, d.j.hand@ic.ac.uk, Department of Mathematics, Imperial College, London, United Kingdom

sequences, we can determine the cause of and possible treatments for disease, but this cannot be done by studying a single gene sequence in isolation. And by studying patterns of credit card transactions, we can detect the account which might be fraudulent; again, this cannot be done by studying a single transaction.

This book, then, describes and illustrates instruments for seeing beyond ourselves, for exploring the properties of objects which we cannot grasp with the unaided brain.

The instruments — tools, methods, techniques — described in this book are very much children of the computer age. To study an aggregate object, described in terms of its individual constituents, requires an ability to sort, extract, combine, and otherwise manipulate the descriptive symbols describing the various attributes of the individuals. Computers provide us with this ability. Computers process the data describing the individuals, converting it into information about them, and then transforming that information into knowledge. This distinction between data, information, and knowledge is an important one. Data are simply symbolic descriptions of the individuals. By themselves they mean nothing. Data with semantics, however, is information. Give me the raw datum that the height of a man is five, and it means nothing. Tell me that the man is five feet tall, and it is useful information. Put it in the context of the general height of men, and it is knowledge, which I can use. Give me a huge body of numerical data and I can do nothing with it. But give me, in addition, the tools illustrated in this book, and I can find relationships, I can recognize structures, and I can detect patterns and anomalies. I can discover knowledge.

The tools illustrated here have a long history. The earliest discipline to concern itself with data analysis was statistics. Since the origins of statistics predate the computer, the aggregate objects with which early statistics dealt necessarily involved relatively few constituent objects. With the advent of the computer, however, the breadth of application of statistics increased. In parallel, other disciplines then began to develop tools for data analysis, typically with slightly different aims and objectives from statistics. Database technology, naturally, was concerned with such problems — not from the perspective of inference, which was always at the base of statistics — but from the perspective of describing and manipulating an existing database. Machine learning appeared on the scene — again, not originally with the aim of analyzing data per se, but rather with the aim of emulating or simulating the way natural systems learnt, and then with the simple aim of building systems which could learn. And, most recently, data mining has appeared in response to the advent of the gigantic data sets, which are now accumulating: data sets of billions of data points, that

is, of billions of constituent objects, are now commonplace. All of these disciplines overlap. They each have valuable lessons to teach each other. A knowledge of one is insufficient without some knowledge of the others. This book demonstrates the application of such tools.
The scope of application of such methods is unlimited. There is no aspect of human life, which is not affected by the need to analyze raw data, by the need to convert data into knowledge. The breadth of different areas discussed in this book demonstrates this beyond question. Furthermore, the exponential increase in the amount of data accumulating, the progress in data acquisition technologies, the dramatic increase in the size of data storage facilities, and the increase in computer power, all of which are discussed in this book, mean that the need for these new tools is becoming ever more important.

I imagine that Galileo and Van Leeuwenhoek must have felt that they were living at the most exciting times in human history, when their tools began to open up the universe to permit the most extraordinary voyages of discovery. The same is true now. The tools described in this book represent a revolution in our ability to see and understand the universe around us. They present us with the means by which to take part in unprecedented adventure.

1.1.3 Background and general trends

Pieter Adriaans[2]*, Insets by Jeroen Meij*

The impact of the computer on our society

If one wants to analyze the impact of the rise of the computer age on our society, one has to look at history. The complexity of the tasks people accomplish in a society gives an indication of the amount of co-ordination they are capable of. Co-ordination of labour requires communication; consequently changes in communication patterns induce changes in the organization of a society. Not only in society as a whole, but also the self-image and intellectual capabilities of individuals are deeply affected by the communication media they grow up with.

One of the most influential events in the history of mankind was, without any doubt, the invention of the art of writing. No longer were people dependent on verbal communication and their own memory. Verbal communication is volatile. It has a limited range, 200 meters at the most. It travels at the speed of sound and an utterance dies out almost directly, leaving the audience with nothing, but their weak and unreliable memories of the events. In a society without writing it is difficult to co-ordinate the activities of large groups of people, it is very hard to maintain a stable legal system. People are dependent on storytellers to preserve their sagas and legends.

Writing

We know reasonably well how cuneiform writing emerged in Mesopotamia. Shepherds had to move their herds from one location to another. Some kind of legal device was necessary to guarantee that the shepherds arrived at their destination with the same number of animals as they started with. To this end a kind of bill of lading was conceived. Each animal was represented by a little clay ball. All the balls were packed in a big clay ball and this was shipped together with the herd. At the destination, the big ball was shattered and the content was matched with the actual number of animals that had arrived. One big disadvantage of this method was of course that nobody could look inside the big ball during the trip, but there was a simple solution: why not scribble some signs on the outside of the ball to denote its contents? Later people found out that it was convenient to have shorthand signs for larger numbers of animals and that one could have different signs for male, female, young and old animals. Not before long people realized that since all the information was written at the outside of the clay ball, it was not necessary to have anything inside anymore. One could equally well leave the contents out, flatten the ball to a kind of slab and only use signs to convey the message: cuneiform writing was born. This example shows how the origin of writing is associated with counting, and with the emergence of

2 Prof dr P.W. Adriaans, pietera@illc.uva.nl, The Universiteit van Amsterdam, Faculty of Mathematics, Computer Science, Physics and Astronomy, Institute for Logic, Language and Computation, Amsterdam, The Netherlands, Senior advisor Perot Systems Netherlands BV

administrative and legal systems. Ancient civilizations like the Sumerians, the Egyptians or the Greeks would never have flourished without these advanced administrative systems based on writing.

Telecommunication

The structure of media used in a society has a deep influence on the social, political and economical organization of that society. Large cultural shifts in history are often associated with revolutions in communication technology. Writing as an externalization of memory is related to the rise of the great ancient civilizations. Book printing, the industrialization of the production of external storage, goes hand in hand with the break down of feudal societies and the rise of individualism. The development of a public press, based on a journal that distributes the same information to large numbers of people on a daily basis, has created the phenomenon of a public opinion that is so important in the concept of a modern democratic state. The invention of the telegraph and later the telephone has increased the speed with which information travels enormously, with dramatic consequences for politics, business and science. In the twentieth century movies and television bring images of wildlife in Australia, Arctic expeditions, riots and terrorist attacks into our living room. A citizen in our modern society has learned more than 90% of the facts he or she knows about the world through the media and not through direct observation.

The computer

There are two essential aspects of the computer that determine its influence on our society: the computer is both a modeling machine and a dynamic storage device.

The computer as a modeling machine has its roots in mathematical research at the beginning of the twentieth century. Researchers such as Gödel, Turing and Church laid down the theoretical backgrounds that enable us to interpret the computer as a universal machine. Any scientific result that can be constructed using a finite number of discrete steps can be calculated on a computer. This result lies at the basis of the success of the computer in science. At the beginning of the twenty-first century a deep knowledge of applied computer science is vital for progress in almost any scientific field: biology, medicine, chemistry, physics and even mathematics itself. Chaos theory could not have been developed without the computer. As a result our whole mathematical conception of reality is changing. Up to 1980 scientists approached reality with idealizations that were in essence directly handed down to us from Euclid. Now, we know that the larger part of the real life systems that we study are non-linear and their behaviour cannot adequately be modeled using traditional linear methods. The human genome project would have been impossible without advanced computing techniques and the analysis of the results is a task, the complexity of which,

is mind boggling. The computer will help us to understand ourselves better and maybe even to redefine ourselves as a species.

The computer as dynamic storage device can be seen as an extension of the classical book. Searching for information in books is cumbersome. One needs indexes, concordances and registers in order to make the text accessible according to its various dimensions. The culmination of this development is the classical encyclopedia. In a computer, text loses its most defining characteristic: the linear organization. Text becomes hypertext, and the book, as a single media entity, dissolves into the multimedia experience of the World Wide Web. There is no single text, no single book. There is an ever-growing patchwork of pieces of information: texts, images, sounds and movies, democratically created by anonymous individuals all over the world, and made accessible by advanced index techniques and search engines. Just as in the case of the invention of writing, this new organization of knowledge influences our society, legally, commercially and artistically. The great impact of the IT revolution on our society has developed in various stages.

In the early days computers were mainly used as a replacement of manual administrative labour. The big mainframes in the sixties and minicomputers in the seventies helped to eliminate dull and uninteresting paperwork, but did not fundamentally change the way companies were organized. Even the high penetration of microcomputers in companies in the eighties had little impact on the structure of those companies. People had a PC on their desk and this became as normal as having a telephone at your disposal. Maybe they were able to access a central company database and had some proprietary e-mail facilities, but that was it. The microcomputer entered the realm of small and medium companies and became a commodity in business environments. Still only a few enthusiasts and hackers had computers in their homes and a few isolated top managers used portable computers.

In the beginning of the nineties the computer revolution gradually began to affect the way companies were organized and in the middle of that decennium the impact on society as a whole became visible. In the nineties a big wave of democratization swept over the computer industry. Not companies, but consumers became the early adopters of the IT market and currently the average desktop worker has a more advanced PC at home then at his desk. These machines were used for word processing and database applications, but the consumer very soon discovered the possibilities of games, music, bulletin board systems and virtual reality. For the first time in history the computer as an interactive device for the storage and manipulation of information, entered our normal every-day life. An important side effect of this development was that a

generation of children grew up in an environment where the computer was as normal as the telephone and the television. It is well-known that human beings learn their first language with a different part of their brain than that which is used for second language acquisition. The brain of a young child adapts easily to new technology. For many parents learning to master a microcomputer was like learning a second language at a later age (a cumbersome and complex process that never leads to completely satisfactory results). Their children accepted the computer as a natural part of their environment. They learned to work with it with the same ease as they learned to read and write. The difference between mastering a skill at the cost of hard labour at a later age and learning it effortlessly in childhood is one of the motors behind cultural and technological development. Generation N creates the natural environment in which generation N+1 grows up. The artifacts of generation N program generation N+1. This is one of the explanations of the emergence of fifteen year old dotcom-millionaires who employed their parents at the turn of the millennium.

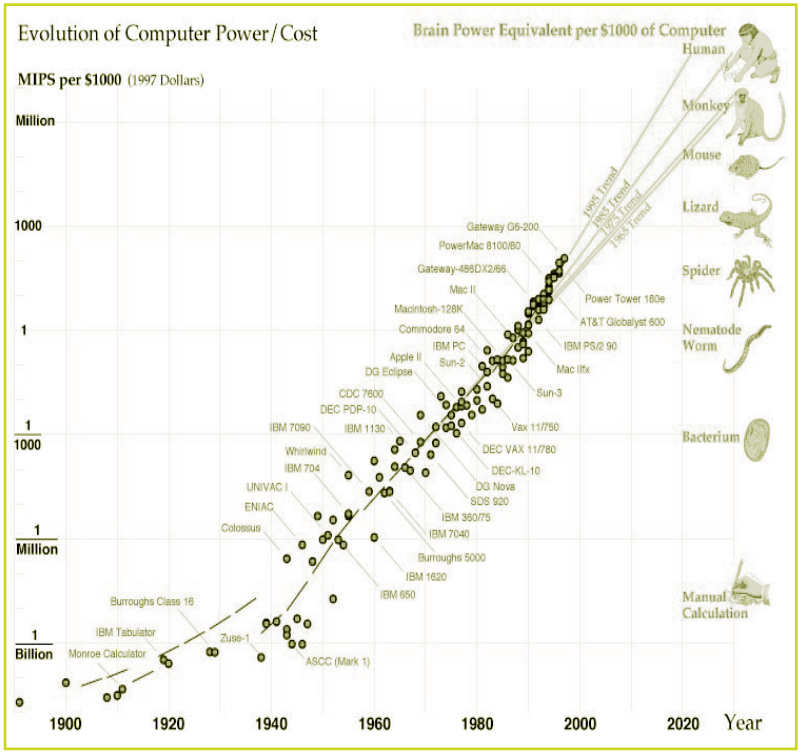


Figure 1
Evolution of computer power/$ 1,000 [Moravec, 1997].

Inset 1: Computing power

Processor power will increase further in the next twenty years, just like data storage capacity. A conservative estimate [Bell, 1997] gives a factor 10,000 increase of processing power (relative to the 1997 level) by 2020. This will enable present 'slow' methods of data analysis to compete with present 'fast' methods in more applications.

The World Wide Web

Looking back it's hard to imagine how difficult it was to connect large groups of users — living in different parts of the world — to one central computer application. Organizations that wanted to do this had to create proprietary local area networks and wide area networks at considerable cost. Only in isolated cases in military environments or in education could the investment in such networks be justified. Users that participated in these networks often had to use dedicated modems for the applications. In such a context a microcomputer on a desk typically would be used for spreadsheets, bookkeeping, word processing, database access and games. The computer user had a one-to-one relationship with his machine and was isolated from the rest of the world. It was only natural that at some point in time the isolated networks would be connected to each other and central data highways, and that a standardized protocol for exchange of information would emerge. This is the birth of the World Wide Web. In a few years hundreds of millions of computers that lived an isolated deaf-mute life on the office desks in New York, hobby rooms in Paris and sheds in Africa were connected to each other, and crystallized into the biggest body of data and information that mankind had ever seen. The Internet revolution was there.

Inset 2: Network society

The (Western) world is heading for the network society, and with great speed. From a life in locally oriented groups to a life in worldwide networks. The pressures and identity that come with belonging to a group decrease, while chances, unforeseen events, globalization and uncertainty increase by participating in social networks. Bases for interactions shift from properties that people are born with (race, sex) to properties that are taken on during one's life (life style, shared values or interests) [Wellmann, 1999].

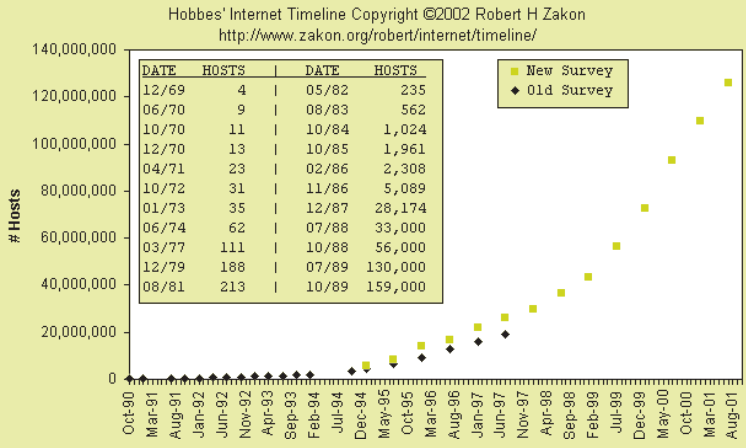


Figure 2
Networks: Internet growth: number of hosts [Zakon, 1999].

The economic and social impact of this development can be compared to that of the invention of the telephone or the art of writing. The nature of the telephone is communication and connectivity. The ability to have a conversation via the phone with somebody who is possibly thousands of miles away is in itself of great value. The laws of graph theory, however, govern the economics of connectivity. They tell us that the economic value of the telephone rises exponentially with the number of people that have access to this medium (i.e. as long as the capacity of the lines is sufficient). The same holds for the Internet with the addition that this medium not only provides connectivity, but also access to data. The organic growth of information and data on Internet gains value every day. The amount of information stored in the world roughly doubles every 6 months.

Inset 3: Network bandwidth

Network bandwidth will increase with a factor between 1,000 and 10,000, depending on the type of network [Bell, 1997]. The 'last mile' remains the bottleneck in the system.

This development has some interesting consequences. Firstly, there is the possibility of doing co-operative work. An architect in Tokyo can design a building in AutoCAD together with colleagues in Amsterdam and New York. Using virtual reality techniques he can give potential clients in San Francisco a tour round the building and make changes to the design on the fly. With a good source-code-control-system, software developers from all over the world can work on one application together 24 hours per day. The results of research in biocomputer science are stored in large database systems that are updated and consulted permanently by large groups of researchers (See 2.2.5, Data access for atmospheric research). The computer gives us possibilities of working together on projects of a complexity previously unknown in history.

Inset 4: Knowledge economy and knowledge society

The value of knowledge increases fast. from a purely economical point of view, but also for the well-being of individual people. Examples of the economical value of knowledge are sales of software or microprocessors. Both are negligible in terms of physical value, compared to the knowledge of programming, respectively digital electronics and manufacturing technology that is represented in them [Stewart, 1997].

Data flood

A simple consequence of the fact that the amount of information stored in computers grows exponentially is that there will be a shortage of people to interpret and understand this information. This holds not only for scientific ventures like the human genome project, but also for more down to earth applications like

yield management in logistics, call detail records in telecom and customer relations management in the banking world. Nowadays a large company may produce more information per day than one individual can digest in a lifetime. This problem is aggravated by the fact that most information in databases is created automatically, without any human intervention or evaluation.

Inset 5: Integration and proliferation of information technology and sensors

The separate network functionalities we know today will be integrated in one network. Equipment for telephone, video, TV and data will converge into a range of combined appliances. Processors (one chip computers) will be integrated in many everyday appliances. Speech interaction with these appliances will be normal by 2010. Our domestic and professional environment will become increasingly 'aware' through numerous sensors that are being incorporated in machines, appliances and everyday objects. This awareness will lead to an intelligent environment, often dubbed as ambient intelligence, or ubiquitous computing.

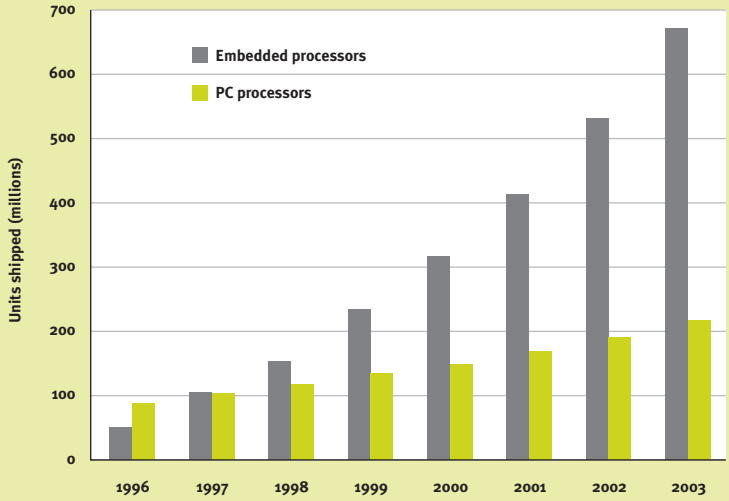


Figure 3
Integration: sales projections for embedded and PC processors [Hennessy, 1999].

When you buy a can of soup in the supermarket, the inventory of the supermarket and your credit card account are updated automatically. You leave the shop; your image is stored in the video surveillance system. You surf on the web; your IP address and the pages you visit are logged automatically in the databases of the marketers. You make a call via the phone and a call-detail record is entered in the system. If the vision of ambient computing of the European community[3] materializes, every individual will create at least a couple of thousands of entries in databases per day in the near future. The world will contain Terabytes of information that no one will ever see.

3 See ISTAC scenarios on the CD-rom.

Inset 6: Digital data

Automated data acquisition and the rise of the Internet are contributing to a strong increase in the amount of electronically available data.

In general a strong tendency exists towards registration and storage of data, often tied to personal data. Moreover, the data is available in a digital form, which has important consequences:

- Digital data can easily be transferred from one carrier to another.
- Digital data can be transported in large quantities without transportation of mass.
- Digital data can be separated from the presentation.
- Digital data can be analyzed by algorithms.
- Digital data can easily be combined with other data.
- Digital data can easily be edited and altered.

This would eventually mean that information could be made available instantly to everyone with access to the live data carrier.

As stated before, currently the access to data is growing fast, as far as the available quantity of data and the amount of people that can access it is concerned.

Data mining

Our society is not very well equipped to handle this abundance of data. For centuries people have worked in an environment that was characterized by a shortage of information. Up to the end of the twentieth century having access to a library, was a vital condition for any form of sensible scientific research. Legend has it that Nathan Rothschild gained a fortune on the London stock market, because he knew the outcome of the battle at Waterloo before anybody else. He used homing pigeons. Today no homing pigeon or even a bulldozer could carry the amount of information that would be available on the Web almost immediately after such events. The emphasis has shifted from gathering data to selecting and filtering data. This development together with the discovery of simple algorithms that work efficiently on large data sets is responsible for the data mining hype in the middle of the nineties. In the mean time data mining with decision trees, association rules and nearest neighbor algorithms has become proven technology.

Inset 7: Software and systems

Applications will drive a cross-fertilization between machine learning, artificial intelligence and many other techniques now used for data mining. Also evolutionary developments in these areas will continue. However, software cost are not expected to decrease in pace with hardware cost, but are likely to remain constant.

Increasingly, knowledge acquisition will be performed not only through, but also on behalf of learning systems, to improve their functionality.

There are lots of vendors who offer sophisticated toolboxes that allow companies to create their own data mining solutions. At the other end of the spectrum we find specialized companies that sell data mining solutions in the CRM[4] market, the area of yield management and fraud detection.

Data mining trends

There are a number of developments in the data mining research that illustrate in what direction the applications are moving. We mention a few.

From batch to on-line. Traditional data mining solutions are batch oriented. Large collections of data are stored in a data warehouse and once a week or once per month a set of data mining algorithms is processed to see, if any interesting patterns emerge. In the case of applications like fraud detection or production control this has serious drawbacks. If there is a flaw in the production process, a company wants to take immediate action. The same holds for fraud detection. There is a tendency to apply data mining techniques directly to production databases. In this case one does not have the benefit of an optimized data warehouse architecture and this calls for new solutions.

From classification to action. Most current data mining algorithms do a great job in classification. They detect that a certain production process is in an error state, or that certain transaction might be fraudulent. In such cases a company would like to take direct automated action. In system management for example one might want to increase the paging space of a certain server or redirect queries to a different data base server. If a system detects credit card fraud, the owner of the card must be warned as soon as possible. This desire leads to the merging of data mining technology with agent technology.

From single table to multi-relational. Data mining algorithms like decision trees and association rules presuppose that the data are stored in a single table. Most data mining applications operate on a single flattened table in which the semantic structure given by the data scheme of the original application is lost. Because of this the mining process is less effective. It might miss patterns that could easily be detected, if the original structure was still available. On the other hand it might find patterns that are trivial results of the process of flattening. Currently research is focusing on creating variants of data mining algorithms that can operate directly on the original set of tables.

From single media to multimedia. Current data mining algorithms work on data that is stored in tables of a relational database, the format of which is heavily restricted. One also would like to be able to mine databases containing images, sounds, speech, music and movies. More advanced techniques need to be developed for this. The combination of mining techniques operating on different media will lead to fascinating new applications, e.g. forensic solutions that mine a database with photos, movies, speech fragments and emails of suspects.

4 Customer Relations Management.

This list shows that the use and scope of data mining technology is still growing. So is its impact on society. The PC revolution was a democratic revolution. Citizens got a new digital freedom. Totalitarian regimes try to keep their subjects away from the Internet, because they know that it is much easier to silence the press than to control propaganda on the Web. Anybody can participate in the making of this enormous database; write new applications, new web sites and distribute information in no time all over the world. This creates not only new possibilities for individuals, but also poses new threats for society. A teenager with a bit of computer experience can easily build a virus that will infect millions of machines all over the world in a couple of days. The networked society seems to be extremely vulnerable to attacks by this kind of hacker. The virus scan applications that are needed to defend us against these intrusions are an example of sophisticated on-line data mining software. If young inexperienced programmers can do such harm, one can imagine what kind of damage a team of seasoned malevolent hackers could do. Cyber terrorism will grow. Future wars will without any doubt also be digital wars. Our society has to take precautions against these threads and data mining technology will play a part in this struggle. The important application areas of this technology, however, will still be science, art and business. The possibility to recognize patterns in extremely large data sets will enable us to explore new horizons. It will help us to create better, more efficient businesses that deliver higher quality in less time, produce fascinating entertainment and new works of art and it will help us to deepen our understanding of nature and of ourselves.

1.1.4 How to read this book

Apart from this Introductory Part, the book does not require reading from the first page to the last. The Parts 2, 3, 4 and 5 can be read in any order desired, though within the chapters there is always an introduction that can assist in understanding the rest of the chapter. Part 6 contains a short introduction to the basics and terminology of the field. The rest of the chapter consists of more detailed technical information that can be used to deepen the knowledge from the other chapters. When appropriate, links to this chapter are provided from the other chapters. More in depth articles can be found on the CD-rom, that also contains a hyper linked version of this book and a text mining tool. A full directory of the CD-rom contents is given in the Appendix.

References

- Bell, C.G., J. Gray. (1997). The Revolution yet to Happen, Microsoft Research Advanced Technology Division. Technical Report MSR-TR-98-44
- Economist. (1999). Drowning in Data. The Economist. June 26
- Hennessy, J. (1999). The Future of Systems Research. Computer. IEEE
- Lawrence, S., C.L. Giles. (1999). Nature. July
- Moravec, H. (1998). When Will Computer Hardware Match the Human Brain? Journal of Transhumanism **1**
- Schreiber, A.Th. (e.a.). (1998). Knowledge Engineering and Management, The CommonKADS Methodology. Version 1.1. University of Amsterdam
- Stewart, T. (1997). Intellectual Capital
- Szalay, A.S. (1999). The Sloan Digital Sky Survey. Computing in Science & Engineering. IEEE. March-April
- Wellman, B. (1999). Living Networked in a Wired World. IEEE Intelligent Systems. January/February
- Zakon, R.H. (1999). Hobbes' Internet Timeline. Version 4.1. April. http://info.isoc.org/guest/zakon/Internet/History/HIT.html

1

1.2
Executive summary

Jeroen Meij

Part 2 **Science**

General

In many fields of science, user-centered systems, standardization and access to data and computing power are recurring needs. Computing (analytic) power will be increasingly *distributed* in what is called a Grid. This will eventually mean that scientists can rent or borrow the necessary computing power and software, when they need it from the Grid.

The analysis of spatial, temporal and relational data is generally considered an important area of development.

Domain knowledge

To advance the availability of online documents, a standardized searchable format should be adopted. XML (eXtensible Markup Language) is a contender for the preferred future format, since it incorporates a strict division between content and shape. Moreover, especially in the case of books, some form of copyright management has to be agreed upon. Since production and distribution costs for digital documents are negligible, digital documents should be substantially cheaper than printed books.

Search

Intelligent search methods will aid scientists in their search for knowledge, meta-search and agent technology being the first steps. Mapping technology will visualize scientific fields and the connections between them.
Interactive visualization techniques for search results will reveal relations between scientific publications and terms, and help in the understanding of related domains.
Since the internet is constantly changing, we can expect scientists to build and manage their own digital libraries, either on-line or off-line, privately or with colleagues.

Scientific methodology

The common hypothesis-testing procedure of science is and will be increasingly complemented with hypothesis generation, especially in areas with large quantities of data.
In many areas, data and interactions between variables are so numerous and complex, that other methods are infeasible.

Expectations for science areas

Life sciences

It will come as no surprise that bioinformatics will play an increasing role in the development and application of drugs. The integration of chemical, biological and clinical data will be a key factor in this development.
Closely related to this, the implementation of electronic patient records and hospital information systems is absolutely necessary for further advancement of knowledge discovery from medical data. The data should not be limited to numerical data, but should also contain results from physical examinations and images or 3D scans. For a rapid advancement of medical knowledge the standardization of hospital data (i.e. by ISO TC215) is imperative, especially in the light of bioinformatics. Hypothesis discovery, temporal diagnostic-pattern discovery, short and long-term therapeutic effects assessment and discovery of new diseases are all expected to follow from medical knowledge discovery in the next decades. Analysis of temporal and relational data can be considered especially important in this field, and both areas will require a lot of development.
Another important issue is the integration of domain knowledge and data derived knowledge. Successful integration will lead to reliable decision support systems that can be used for education and to assist with diagnosis for experts and patients.

Environmental and ecological sciences

GBIF, the Global Biodiversity Information Facility of the OECD, could provide a

much needed unification of separate databases covering museum collections of organisms all over the world. A convergence of bioinformatics and biodiversity research is expected.
For the environmental sciences, a need can be identified for better data access and preparation, supported by distributed computing power. Special attention should be given to spatial and time-series analysis, from data representation to knowledge representation. A strong need for standards for spatio-temporal models can be observed, which are expected to lead to implementations in current GIS and database systems.

Economics

We can train agents based on historical data. With these agents we can simulate social processes through emergent behavior, thus creating adaptive social simulation systems, useful for practical and fundamental research purposes. We could derive market behavior, observe social trends and market mechanisms and simulate the consequences of political measures.

Keywords

Summarizing the data mining related developments for science in a few keywords, we expect to see:

- integration of functionality and fusion of databases;
- agents assisting in acquiring domain knowledge;
- integration of data-derived knowledge and domain knowledge;
- distribution of data and computing power.

However, some obstacles will require our attention:

- restrictions on access to domain knowledge (standards and intellectual property);
- lack of standardization of data formats within domains;
- poor user friendliness of systems.

As the trends mentioned above become reality — provided these obstacles are negotiated — science will advance faster then ever before.

Part 3 Business and government use

General

For business users, government users and consumers alike, data mining is expected to move from the domain of the specialist (data miner, statistician, scientist) into the domain of the user. First reaching the business analysts and marketeers, then reaching business and private customers. The democratiza-

tion of data mining as a process is just starting and will continue. As with many other successful developments, data mining tools will be embedded in a wide range of software products and services, often without the end-user realizing it. With progressing virtualization of business (see below), the need for data mining tools grows, and this will ultimately lead to real time contextual adaptation.

Customer relations management

It seems reasonable to assume that the trend towards more personalized and ultimately one-to-one marketing will continue. Even so, one-to-one relationships must be meaningful from the perspective of the customers as well. Ultimately, good data mining practice — involving ethical as well as commercial principles[1] — could lead to benefits for both the selling companies and the customers: less undesired sales contacts, and a higher percentage of welcomed sales contacts.

Customer emancipation is imminent: when the appropriate software tools (or services) are available, intelligent agents will scour electronic markets, representing individuals or groups of customers to search for necessary, interesting and useful products. These customers will only release profile information, when they feel it is to their benefit.

Tasks

There are several identifiable business tasks that can be related to data mining actions.

Grouping (clustering, segmenting)

What distinct groups exist within my customer base?

Being able to discern a group of entities with common characteristics. From a mass of data one or more useful groups are identified. Examples might be customer groups or demographic regions.

Data fusion methods could allow us to enrich entire customer databases with survey information that is only available for a sample, in other words, carrying out a virtual survey with each customer.

Categorization (classification)

Which group does this new customer belong to?

Assigning an entity to a known category. Examples might be assigning customers to a known group, separating different quality classes of fruit, separating different vehicle types.

Developing automated adaptation systems will be a logical next step. Systems such as these are expected to be used in many areas, quality inspection being an area of particular interest.

1 See Chapter 4.1

Detecting

If I only had a warning light, indicating we should investigate these particular cases..

Being able to detect a deviating state that can be considered relevant. Intrusion detection in a network, phone or money-transfer patterns that indicate fraud can be seen as examples.

As data collection increases, so does the importance of automated detection. In many cases it is sufficient to know when human interference is required, and automated detection is a cheap alternative to human observation and analysis. An interesting question for the near future is whether people prefer to be monitored by humans or by computer systems.

Modeling

What are the influences of family size, income and ..? on choosing a car?

Generating an abstract description of (a part of) reality. This mainly concerns the cases where we are interested primarily in understanding processes. With this understanding we can develop better regulating mechanisms or products.

Prediction

What car models will this man be interested in?

Predicting behavior of groups, individuals or systems. When we have a model, we can also predict the outcome for new values, provided the process itself does not change. We can predict which clients will be interested in a caravan policy, or will be paying their credit card debts in time.

Matching

Which offers are able to fulfill this request?

Creating a useful link between two or more entities. Examples might be job-matching, purchase/sales matching or dating.

We will see data mining — combined with agent technology — playing an important role in many matching and linking situations. This applies to business to business, but also to business to consumer and consumer–consumer relations.

Adapting

What should our homepage look like, when we are visited by a sports enthusiast? and when a teenager visits the page?

Being able to adapt a system to a situation (or customer). Examples might be web pages, educational systems and procedures.

We can expect self adapting systems for many of the tasks described in this book. These systems will adapt themselves to new conditions, products, quality demands, etc. of a process.

Keywords

To make a very condensed summary, we expect the following data mining related trends to materialize in business and government:

- democratization of data mining;
- integration of data mining in many business processes;
- automation of adaptation cycles;
- agents aiding the emancipation of consumers.

Privacy, ethical and legal aspects need our attention.

Part 4 Ethical and legal aspects

Ethical perspectives on web mining

Web mining technology is already being used for many commercial purposes. Generally, web miners benefit most from web mining, while web users are facing the dangers.

Although the impact of web mining should be of every web user's concern, there is no reason to panic:

- the technique is not yet being used to its full potential;
- there is no clear indication of web data being misused to such an extent that people are hurt by it.

One of the dangers lies in the hidden way in which web mining can be used. Companies can cover up their ultimate goals, when they obtain certain bits of information. As web mining is in an early stage of development, there are things that can — and need to — be done to guide this technique in a socially acceptable direction. Since ethical issues will grow as rapidly as the technology, ethical considerations should be an integrated and essential part of this development process. Since no ethical guidelines can cover every possible misuse, we need to realize the seriousness of the dangers and to continuously discuss the ethical issues. This is a joint responsibility for web miners (both adopters and developers), web users and governments.

Legal aspects of data mining

Fair information practices

The OECD formulated eight principles, called the fair information practices, which may be used to evaluate data collection and processing. These concern collection limitation, data quality, purpose specification, use limitation, security safeguards, openness, individual participation and accountability. These principles have been implemented in legislation in many countries around the world.

Legitimacy of decision rules

Decision rules should be well founded, usable, allowed and acceptable. This means the rule should have a proper motivation, for both its criteria and the decision in the decision rule.

Regulation of law enforcement

Data mining by law enforcement authorities is generally regulated through special legislation superseding general privacy laws. Nevertheless, the data should be acquired, processed and stored legitimately. Especially data mining not directed to a specific crime or suspect is bound by more strict regulations.

Part 5 The perspective of the individual

Data and data conservation

From technical developments in applications alone, we may expect a large increase in surveillance data, web mining data and survey data about individuals. Some of these applications will have to be subject to political debate to determine the actual level of proliferation desired.
The same activities will also be responsible for the availability of large quantities of data for individual use. People will create their own digital collections of literature, web pages, articles, music, pictures and video.

Individuals or organizations depending on specific information need guaranteed access. This would mean either a public (national, global) Internet archive, or the creation of a private or local archive that contains important information.

Multimedia mining

In the next decades, we can expect real multimedia mining applications to enter the commercial realm. This will be made possible through the development of dedicated algorithms and the close collaboration between data, sound, video, image processing experts.
In the very near future, we will see multimedia data mining tools as applications in cars, in homes, and even with wearables (i.e. computer powered cameras built into garments). Cameras mounted on computer displays could identify user emotions and interpret needs. Identifying and recognizing objects in real time will become common practice. Cameras mounted on mobile carriers, such as cars or even humans, will have enough computing power to help users recognize and interpret the environment in which they proceed. Such devices could help car drivers in tracking potentially dangerous situations or warning the driver of fatigue or distraction.

Image

First, large scale image databases are being created.
Second, research is directed towards the integration of different information modalities such as text, pictorial, and motion. Third, relevance feedback will be and still is an important issue. Indexing, searching and assessing the content of large scale image databases will be done by software tools, not by humans.

Video

Product suites for content-based image and video indexing and searching will be developed. These tools will serve the needs of future content owners in the field of entertainment, news, education, video communication and distribution.

Music

In the next stages of development, musical audio mining products will be employed by professional content distributors, entertainment and leisure industry and, finally, by the consumer.

Text mining

Combining text mining and data mining technology with general machine learning technology will yield a next generation of intelligent adaptive knowledge management systems. These knowledge management systems will be able to increase their knowledge of the domain with the growing number of documents contained in the system. Moreover, the adaptive knowledge management system will be able to adjust its knowledge, when documents from new domains are added to the system. The next generation knowledge management systems will be of particular interest for multidisciplinary and fast changing markets, such as the professional services organization industry.

Web mining

There is cross-fertilization between information retrieval and extraction on the one hand, and data mining on the other hand. Both may be useful as a component of the other. On the assumption that the current trend continues, it is reasonable to expect that in the next decade the Web will evolve into a knowledge base, the completeness and intelligence of which will largely surpass that of any encyclopedia, newspaper or classical library, and, for many domains, even that of human experts.

Knowledge integration and learning

The use of personalized knowledge profiles, which describe ‘gaps’ in the knowledge domain of an individual, will be an important step forward in the life long learning perspective of the knowledge worker. For best results, a way should be found to determine the potentialities of both competencies and interests of a

person, and based on that to assist him or her in data mining. But even then, ample room should be left for individual choices enabling self-tuition.

Future knowledge workers

Now, can we envision the changes that will occur when mature video, audio, text and multimedia mining tools are commercially available? Where agents know our behavioral patterns and will react on what we experience, and anticipate on what we want?

From the developments sketched in this part, we can formulate a vision of future knowledge workers. When we combine such a general vision with three different profiles of knowledge workers, we might see different levels of intensity of use.
The first group of people are permanently connected to the Internet, but also have their own data repository. They are assisted in performing their tasks by advanced search, analysis and presentation (summarization, graphics) software. Software agents continue gathering, while they are doing other things. Input is mainly text (typed or voice) or graphic interface based, output on a screen or head mounted display.
Interacting in a more intense way, another group of people will be immersed in their search and productive environment at times, where interaction with advanced tactile and movement sensors and actuators enables them to explore and act. These immersion techniques will also make virtual presence and cooperation possible.
The highest interaction intensity will be reached by those who will be connected through clothing, accessories and implants during a large part of their day, working in a highly augmented reality. Context aware agents supply them with additional information on their physical or search environment continually. Besides having the advantage of being well informed at all times, they probably will have short periods of absentmindedness, when they are communicating with the system or with each other.

Part 6 Methods and technology

Definition

We propose the following definition of data mining:

Data mining is the process of extracting previously unknown information from aggregations of data. In the right context, this leads to knowledge.

Usually, the data mining step is embedded in a larger process, the Knowledge

Discovery (KDD) process. We can make a division of the KDD process in the following steps:

- Problem analysis.
- Data acquisition.
- Data processing.
- Data analysis.
- Reporting.

In turn, the KDD process is embedded in the business process. Last but not least, technical embedding in the IT environment is an important issue.

For any business task (and question) described in Part 3, many paths and techniques are available to resolve the task and answer the question. Choice of the right path and technique is a matter of expert judgment, although supportive tools are being developed.

Several technical trends can be observed in the area of data mining.

Hardware

In hardware there is a distinct trend towards distributed and parallel computing. As I/O and operating systems improve, ccNUMA machines will play an important role in data mining applications. In the near future, the application of cheap Beowulf clusters development (of workstations or PC 's) will rise as a result of increasing network speed, as for example defined in the Infiniband protocol. Field Programmable Gate Arrays (FPGAs) hardware may be configured to perform new tasks, achieving the optimal configuration for every operation.

Parallel data mining

Parallel execution of different data mining algorithms and techniques can be integrated to obtain a better model, not just to get high performance, but also high accuracy. These techniques may lead to environments and tools for interactive high performance data mining and knowledge discovery, including parallel text mining, parallel and distributed web mining. Other interesting developments are the integration of parallel data mining with parallel data warehouses, and the integrated use of clusters and grids for distributed and parallel knowledge discovery.

Relational data mining

A new development is data mining on relational data, which is being extended to object oriented databases. Domain knowledge and distributed environments can be integrated in the data mining process by using the object oriented UML, Unified Modeling Language.

Pattern evolution

When data mining has revealed patterns from a database, an evolution in the patterns will occur when the data changes. A framework that is capable of dealing with all changes a rule may undergo, is still missing, and might be a direction for future work.

2

Knowledge Discovery in Science

2.1 Introduction

Jeroen Meij

Many sciences have to deal with large amounts of data gathered from observations. Automated data collection has opened further possibilities for gathering data, often collecting far more data than is actually needed for one experiment. It is also becoming increasingly common to collect large quantities of data of which the greater part contains no useful or interesting information, although the scientist is only interested in the observations that are 'out of the ordinary'. Needless to say, in areas like astronomy, biology and environmental sciences the demand for intelligent data analysis tools is high.

At the other end of the scale, we find the social sciences, in which data quantities are usually small. A typical problem in this area might be the imbalance between the number of variables and the sample size. Data mining methods can be used to select the variables that will be most suitable for analysis.
This part of the book takes the scientist's point of view, looking specifically at methods that generate knowledge from data. We will focus on three theme's that are subject to change through the increased use of information technology: domain knowledge, access to data, and scientific methodology. The chapter concludes with examples from different sciences.

Domain knowledge

First, we try to sketch some approaches for the benefit of any person seeking scientific information or conducting scientific research. On one hand, this includes gaining insight into the position and coverage of the scientific field itself, related to neighboring fields and evolving in time. On the other hand the scientist's knowledge is enriched and updated by access to scientific publications, made more accessible through text mining tools or with the help of information agents. The central role of domain knowledge in the scientific process is illustrated in Figure 1 and 2.

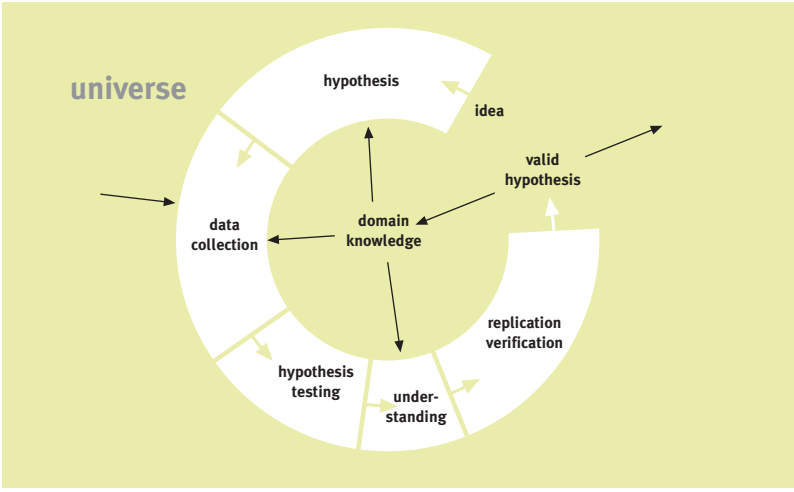


Figure 1
Traditional scientific process.

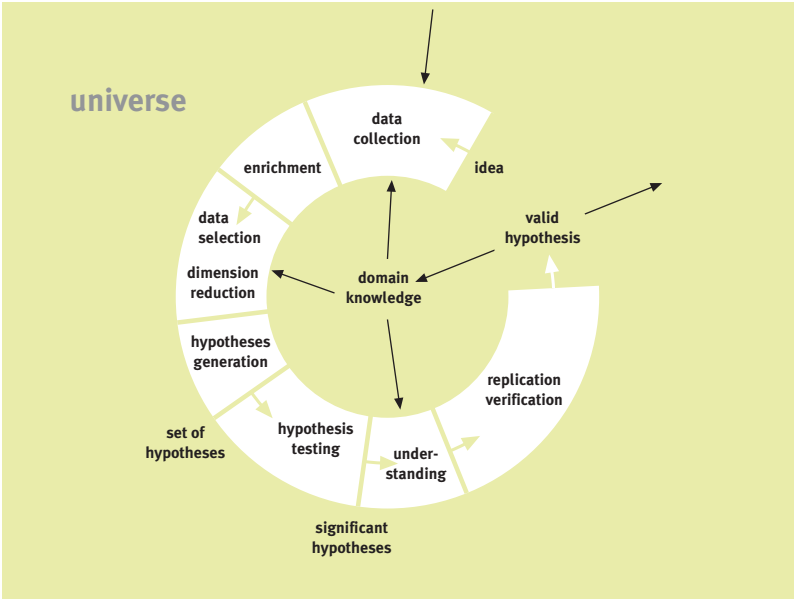
Data collection and access

The use of data gathered by others and available through various distributed sources is becoming standard practice (see Section 2.2.5, Data access for atmospheric research). This requires new technical infrastructure, but also new standards and interfaces.

Mining for hypotheses

Besides the facilitation of access to domain knowledge and data, there is a trend towards the recreation of the core of the scientific process itself. Exploratory data analysis was once regarded merely as a useful tool to describe variation in a data set. Combined with elements from hypothesis driven analysis, however, a new methodology is emerging. The focus is shifting from 'starting with a good hypothesis' towards 'starting with good data' [Clancey, 1984; Palmer, 2001; Adriaans,1998]. Nevertheless, it should be stressed, that good statistical practice should not be neglected. This matter is discussed in Section 2.2.4, Mining for scientific hypotheses.
An impression of the influence of data mining on scientific practice in general is given in Figure 2.

Figure 2
Scientific process using data mining techniques.



Examples of data mining in scientific fields

The remainder of part 2 consists of cases and prospects from many scientific fields. Each case is written by scientists from the specific field, who uses data mining tools in his or her work. Application oriented fields include biosciences, medical science, linguistics, economics, environmental and behavioral sciences. Many of them give views from — at first sight — unexpected angles. We will conclude with a general outlook on the role of data mining in science in the next decades. For a detailed discussion of the techniques mentioned in the cases we refer to Part 6 of this book.

References

- Clancey, W.J., E.H. Shortliffe. (eds.). (1984). Readings in Medical Artificial Intelligence, the First Decade. Addison-Wesley, London
- Palmer, M. (2001). Hypothesis-Driven and Exploratory Data Analysis. http://okstate.edu/artsci/botany/ordinate/motivate.htm
- Adriaans, P., D. Zantinge. (1998). Data Mining, Addison-Wesley, London

2

2.2 General Application for Science

2.2.1 Text mining for science

Antal van den Bosch[1]

Introduction

Text mining (or text data mining) is a phrase coined in the late 1990s to denote the discovery of knowledge in texts. The underlying idea is that natural-language text is one of the most-used means of conveying or storing information other than through spoken language; in many real-world domains large sources of material are available as printed and electronically stored text. Especially in knowledge-rich contexts, such as science, texts play a vital role in communicating, conveying information and establishing new facts.

From a data mining point of view, the bottleneck in automating text mining is the natural language in which the texts are coded. Natural language has evolved through many centuries into a complex code system; each world language is in itself essentially cryptic to outsiders without access to the model (or rather, the morpho-phonological, syntactic, semantic, and pragmatic sub models) by which that language can be analyzed. In mining terms, the soil through which the text miner needs to drill in order to arrive at valuable information and knowledge, consists of a complex system of layers which do not allow a direct bore, but rather a costly, stepwise drilling procedure involving many different techniques. Designing this procedure is a common goal in the fields of computational linguistics and of information retrieval, and the problem is presently far from solved [Hearst, 1999].

1 Dr A. van den Bosch, Antal.vdnBosch@kub.nl, ILK / Computational Linguistics, Tilburg University, The Netherlands, http://ilk.kub.nl/~antalb/

In this contribution we first identify the possibilities and difficulties of mining knowledge from scientific texts. We then give two examples of text mining in scientific domains (medicine and computer science). The section closes with a summary of the prospects in this area.

State of the art in text mining for science

Since the late 1950s there have been attempts to model the process of scientific discovery on computers [Newell, 1962; Langley, 1987]. The methods produced in this strand of artificial intelligence tend to avoid the use of natural language; they abstract from it in order to have models of scientific discovery that are not 'contaminated' by the complexity and noisiness of the language model.

With the development in the 1990s of better language modeling methods in computational linguistics, largely based on brute-force computational methods drawing on large electronic data bases of sampled natural language [Manning, 1999], the notion arose that scientific texts could be mined directly to produce knowledge. Although not all of the natural language in a text can be analyzed, a sketchy but certain analysis may often be enough to retrieve the most relevant knowledge from a text. Moreover, language analysis methods generally work better when the texts conform to a certain regular form, and when they are in one domain. And this is what one generally finds in scientific domains.

In all sciences there is at least a tendency to have high standards for published texts. In some domains, for example in the medical or the empirical social sciences, scientific texts can be expected to conform to a standard template. This template forces hypotheses, results, and discussion issues into predictable places, thus making access to the information in the text easy for those who know the template. Even in scientific areas in which texts are more free-form, scientific texts, provided they are written well, usually achieve clarity in presenting the information and knowledge through the use of relatively widely accepted typographical conventions: titles, subtitles, keyword lists, abstracts, the way texts are split in titled sections, subsections, and paragraphs, captions of tables and figures. All this structure can make it considerably easier to extract knowledge from text. This high degree of organization is an obvious advantage for automatic text mining.

Examples

A problematic aspect of working in science is keeping up with new developments. This is so not only because reports of new developments are abundant, but also because they appear in different media which are not all immediately accessible and searchable by one individual. At the same time, it is commonly believed that having such an overview would enable individuals to discover new

facts or theories, simply by combining the components which have been found already, but have not been combined yet. In principle, text mining techniques could be employed to do at least a part of the work of combining such ingredients.

A particular demonstration of this has been provided in the work of Don Swanson and co-workers [Swanson, 1994; Swanson, 1997]. Looking only at titles of medical papers, they have shown how chains of causal implications from these titles can lead to the discovery of hypotheses for the causes of diseases. One example, the causes of migraine headaches, has received supporting experimental evidence. Swanson et al. first cleaned up titles from articles which shared a subset of title words (titles of medical papers tend to be quite clear, but they can contain modal words and sub sentences that can be deleted without harming the global message), and then worked on the following types of 'clues':

- Stress is associated with migraine.
- Stress can lead to loss of magnesium.
- Calcium channel blockers prevent some migraines.
- Magnesium is a natural calcium blocker.
- Spreading cortical depression (SCD) is implicated in some migraines.
- High levels of magnesium inhibit SCD.
- Migraine patients have high platelet aggregrability.
- Magnesium can suppress platelet aggregability.

The apparent implication that magnesium deficiency might play a role in some kinds of migraine headaches (which can be deduced from the above list with logical formulae emulating simple natural language) was not acknowledged as a fact nor as a hypothesis in medicine at the time of Swanson et al.'s experiments in 1987; in fact, a medical paper in 1989 corroborated Swanson's result [Ramadan, 1989].

A second example that exemplifies the use of text mining in science comes from Carnegie Mellon University's Text Learning Group [2] headed by professor Tom Mitchell. The Web->KB project of this group aims to develop methods to mine the World Wide Web, the largest text corpus that can be accessed freely (as such, web mining is closely related to text mining). As a first example of the methods developed in their group, they developed a system that mines the web to create a taxonomy of courses, lecturers, departments, and students in computer science. This taxonomy can then be used to find information about the domain of university-level education in computer science: unlike scientific discovery, this application extracts structured knowledge automatically from text; knowledge which can be very valuable for anyone working or studying in the area.

2 Homepage of the Text Learning Group of Carnegie Mellon University: http://www.cs.cmu.edu/~TextLearning/

The prototype Web->KB system takes an ontology defining the basic relations in the domain as input (a course has a lecturer; a lecturer is affiliated to a university; etc.), and web pages that are hand-labeled instances of course, lecturer, department, and student home pages. Based on this hand-labeled training material, the system learns to extract information from new pages that can be added straight into the growing knowledge base. New pages are classified into one of the four categories by smart information retrieval that makes use of automatically learned keywords and key phrases (which can include natural language as well as HTML code). The prototype system was able to learn a knowledge base by training on the CMU computer science department web pages with web page classification accuracies in the range of 70%, yielding a relatively accurate searchable structured knowledge base representing the department's entire curriculum.

Prospects

In an overview paper on text mining, [Hearst, 1999] predicts that the best results of future text mining applications will be obtained by the intelligent combination of methods from computational linguistics for (partial) analysis of natural language, and user-guided analysis of automatic text mining results. For science, this would mean that text mining would be most effective embedded in a mining procedure that, for a particular scientific domain, involves both an analysis of texts tailored to the types of text generally occurring in that domain, and an interactive procedure in which experts inspect and enrich the rough outcomes of the natural language analysis stage (which, for the next decade, can be expected to provide sub-optimal performance and accuracy, as long as the text is natural).

Text mining is on the agenda of several main research groups in the world (e.g. Berkeley's School of Information Management and Systems[3]; CMU's Center for Automated Learning and Discovery[4]), and it is becoming one of the key challenges in computational linguistics and information retrieval, in fact joining the two fields. As much of the leading research is partly motivated by commercial applicability, text mining is also becoming an industry, being the focus business of several new companies. It can be expected that domain-tailored text mining tools (e.g. for medicine) and generic tools (most probably for internet searching) will be available. For science, both developments hold great promise, as it is the feeling, supported by the experience that one gets by surfing the internet 'manually', that vast amounts of information and knowledge are still untapped, and most importantly, that ingredients for new knowledge are available and their vital combination can be discovered with text mining.

3 Homepage of the School of Information Management and Systems, Berkeley: http://www.sims.berkeley.edu

4 Homepage of the Center for Automated Learning and Discovery, CMU: http://www.cs.cmu.edu/~cald/

References

- Hearst, M. (1999). Untangling Text Mining, in: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. ACL, New Brunswick, NY
- Langley, P., H. Simon, G. Bradshaw, J. Zytkow. (1987). Scientific Discovery: Computational Explorations of the Creative Processes. The MIT Press, Boston, MA
- Newell, A., J.C. Shaw, H. Simon. (1962). The Process of Creative Thinking. In: H.E. Gruber et al. (eds.). Contemporary Approaches to Creative Thinking. Atherton, New York, NY
- Ramadan, N.M., H. Halvorson, A. Vandelinde, S.R. Levine. Low Brain Magnesium in Migraine. Headache **29** (7):416-419
- Swanson, D., N.R. Smalheiser. (1994). Assessing a Gap in the Biomedical Literature: Magnesium Deficiency and Neurologic Disease. Neuroscience Research Communications **15**:1-9
- Swanson, D., N.R. Smalheiser. (1997). An Interactive System for Finding Complementary Literatures: a Stimulus to Scientific Discovery. Artificial Intelligence **91**:183-203

2.2.2 Agents serving science

Wiebe van der Hoek[1]

Abstract

How do scientists find what they are looking for? The quantity of information available, the number of sources and the ease with which it can be obtained have changed dramatically, as a result of new technology. Conventional search methods are no longer optimal, if applicable at all. We briefly examine the paradigm of intelligent agents, and examine whether and how they can help the scientist to find and manage the information available.

Introduction

In a letter to Hooke, written 1676, Sir Isaac Newton wrote "If I have seen farther than others, it is because I have stood on the shoulders of giants" (Turnbull, 1959, taken from [Harmsze, 2000]). It is only because scientists are educated by teachers, and that they debate with out- and insiders in their field of expertise and propound their problems and results to a platform of colleagues, that science is able to continuously shift its frontiers.

Having ascertained that exchange of information is vital for scientists, we further diagnose that radical changes are taking place in the area of information supply and demand. First of all, the form in which information is available has changed drastically. Also, the quantity in which information is available, the number of sources and the ease with which it can be obtained change dramatically. In short time, hundreds of millions of people have fast, pervasive access to a phenomenal amount of information, through desktop machines in the office, school and home, through televisions, phones, pagers, and car dashboards, from anywhere and at anytime. Conventional search methods are unable to deal with these changes. As [Hermans, 1997] puts it, these methods are based on the principle that it is known which information is available and where exactly it can be found. However, due to the size, the format of documents and the dynamic nature of the Internet, such methods are no longer optimal, if applicable at all. The following quote from [Naisbett, 1980] summarizes this nicely: "We are drowning in information but starved of knowledge."
In this paper, we describe some technologies that aim to assist the user and in particular the scientist, in finding and managing the information that becomes (electronically) available. We focus on the technology offered by so-called *agents*, which, allegedly, show at least a minimum of intelligence in doing their task. We will also describe some existing systems that are based on this idea. The rest of this paper is organized as follows. We stress once more the importance of communication for scientists, and we describe how the channels for

1 Dr W. van der Hoek, wiebe@cs.uu.nl, Department of Computer Science/Department of Philosophy, Utrecht University, The Netherlands, Department of Computer Science, University of Liverpool, The United Kingdom

this are dramatically changing. In the paragraph concerning agents we explain what the agent paradigm stands for, and in the next paragraph we specialize this to information agents. There, we describe some solutions for particular problems, and we introduce some implemented agent-based information systems. In the last paragraph we conclude.

Before starting off, I would like to spend one word on the methodology used in writing this paper. In order to treat the new information age as a serious challenge, I have been working on this article without any paper on my desk. Although I have used one orthodox retrieval method — being a worker in the field of agents, I wrote the paragraph on agents mostly 'from my head' — all the other information reported here was found by electronic means, using some valuable resources on the Internet or CD-rom. In particular, the paragraph on communication for scientists is very much inspired by a CD-rom accompanying the Ph.D thesis of [Harmsze, 2000]. The paragraph on information agents is influenced by a paper by [Klusch, 2001] that is currently only available in electronic form, and also by a paper of [Hermans, 1997], retrieved electronically. Using such rich sources invites to sometimes shameless literal citing, for which the owners hopefully forgive me. I would like to thank the three authors mentioned for their valuable documents, as well as all the other consulted sources (all mentioned in the references).

Communication for scientists

Francis Crick, who won a Nobel prize for the discovery of the molecular structure of DNA, formulated the importance of communication for researchers very accurately: "Communication is the essence of science." This statement of Crick's is quoted in (and even reflected in the title of) [Garvey, 1979]. The types of communication (-channels) that scientists usually employ range from *informal* (like discussion between colleagues) to *formal* (e.g. scientific journal) with many intermediate forms (like preprints). It appears that the rise of computers, and in particular, the Internet, has heavily affected all the types of communication among researchers, within the scientific community and between scientists and the rest of the world. Scientists interchange half ideas and completed papers by e-mail, and subscribe to discussion groups on specialized topics, the Internet is becoming *the* platform to announce conferences or distribute calls for papers, and the immense amount of data stored over the net has become an ever expanding source of information.

Scientific journals

As was also noted in [Harmsze, 2000], the emergence and evolution of the scientific journal, and the process of scientific communication by means of journals, have been prompted by the needs of scientists and by the possibilities

offered by the publication media (for a detailed history of the scientific journal, see the work of [Meadows, 1974; Meadows, 1998]). The first scientific journals were established in the second half of the seventeenth century; until then, small groups of scholars communicated with each other via private correspondence. Harmsze nicely describes how Henry Oldenburg in 1665, as a secretary of the Royal Society, read relevant letters aloud at the meetings of this society. Soon, however, the correspondence started to overwhelm him. He tackled this problem by printing and distributing the most important letters; thus the journal 'Philosophical Transactions: Giving Some Account of the Present Undertakings, Studies and Labours of the Ingenious in Many Considerable Parts of the World' was established. Of course, this development was only made possible by the new service of printing press.

It is interesting to note that the 'Neues medicinisches Wochenblatt fur Aerzte' already in 1789 complained (see [Meadows, 1974, p. 72]): "This is truly the decade of the journal, and one should seek to limit their number rather than to increase them, since there can also be too many periodicals." However, the amount of information communicated via journals has only been growing: [Meadows, 1998] estimates that the number of titles doubles every 10 to 15 years, more articles are published in each journal and the articles get longer. So, almost with the introduction of the printed scientific journal, a significant factor in the shaping of scientific journals is the continuing endeavor to protect scientists from being drowned by the information flow.

Since the Second World War, this problem has become acute. For the individual scientist, it is impossible to keep up with the papers in officially refereed journals, let alone the publications that appear as abstracts, proceedings or notes that are available in a pre-mature state. Moreover, the scientist not only has to keep up with current issues of journals, but he also has to have knowledge of the accumulated archives. The following numbers, that Harmsze finds in [IEEE, 1999] are illustrative: ... "the Institution of Electrical Engineers scans over 4,000 scientific and technical journals and some 2,000 conference publications for the bibliographic INSPEC database. At the end of 1997, the Database contained nearly 6 million bibliographic records and is growing at the rate of 330,000 records a year." As a consequence, on the one hand, there is a massive amount of information available, but, on the other hand, it does not reach the scientists who need the information. In the words of [Garvey, 1979]: "In some disciplines it is occasionally easier to repeat an experiment than it is to determine that the experiment has already been done." This would mean that the publishing of scientific results undermines its own main purpose.

Electronic media

Needless to say, in the scientific work office also, computers have become a major tool (after all, it was in this context that the Internet was created. For a history of the Internet, see [Leiner, 1998]). In a period of 20 years, the typewriter has completely been replaced by the word processor, so that sources are immediately available in electronic form. With the development of networking technology, researchers can directly interchange documents without delay. It is also possible to have a discussion with a colleague, an editor or a reviewer on the other side of the world and exchange ideas and arguments in text, graphics, pictures, sound and databases. New facilities have been specifically offered by the development of the World Wide Web (WWW), initiated in 1990 at the Centre Europeen de Recherches Nucleaires (CERN). Tim Berners-Lee, the inventor of the Web, describes it as a 'distributed heterogeneous collaborative multimedia information system' [Berners-Lee, 1991]. Computer files are distributed via the Internet and stored in on-line electronic databases, which can be retrieved from the net. Sites emerge on specialized topics, in which for instance indexes ('links') to resources are gathered, and which have bulletin boards and notification services, conference lists with deadlines and job openings, and a forum for discussion.

At the same time, scientific journals appear on the net. [Harmsze, 2000] mentions that around 1995 a number of 83 of them in the domain of science, technology and medicine were examined. Now, publishers have electronic versions of almost every journal, and even specific journals for the web are raised. The advantages of such journals are the easy access, and the short time between submission and publication. Also, the technology of the net offers new facilities and opportunities, for instance by adding explicit links to papers. In the subsection on improving the supply of information we will see that Harmsze argues that the idea of 'a paper' should even be completely re-evaluated given the new facilities on the Internet.

Agents

As computer programs and applications tend to become more and more complex, there is a need to move to a more abstract level for designing, implementing and reasoning about novel software. The need to systematically deal with such complex systems was recognized from the early days in computer science, and dealt with by encouraging structured programming, using procedures and introducing the notion of *objects*. However, in the last 10 years the notion of *agents* seems a promising one to organize complex, often distributed, systems, with many interactions. Agents are a logical consequence of (self) organizing software, if one recognizes that even users contribute human-like properties to their computer (applications): "my mail-program *believes* that this message is

not acknowledged", "my word processor *wishes* to open a second file", "the computer *thinks* the file is not available."

Although Hofstadter does not use the word agent explicitly, the same idea must have been in his mind when one of his characters, Sandy, puts the following forward in a Coffee House Conversation [Hofstadter, 1981]: But eventually, when you put enough emotionless calculations together in a huge coordinated organization, you'll get something that has properties on another level. You can see it — in fact you *have* to see it — not as a bunch of little calculations, but as a system of tendencies and desires and beliefs and so on. When things get complicated enough, you're forced to change your level of description. To some extend that's already happening, which is why we use words such as 'want', 'think', 'try', and 'hope' to describe chess programs and other attempts at mechanical thought.

Rational agents are the central objects of study in many research disciplines, including economics [Neuman, 1994], philosophy [Dennet, 1987], cognitive science [Stich, 1983], artificial intelligence and computer science [Wooldridge, 1999]. Agents are then pieces of software, that act in some environment in order to modify, shape or control it, on behalf of their user. Whereas natural examples of agents in computer science and artificial intelligence are provided by physical entities like (autonomous) robots, another example is given by so-called *softbots:* pieces of software that assist the user in a virtual environment. Examples of such softbots vary from small applications that helps the user mastering his word processor, personal assistants that schedule meetings and maintain an agenda, or information agents that represent the user on the Web and help him find relevant information. Although a precise definition of agents is lacking, pioneers in the field like Wooldridge and Jenings [Wooldridge, 1995] distinguish the following properties of agents:

- *Autonomy*. Whereas objects can be invoked by other objects, software agents decide by themselves what their next action will be.
- *Reactiveness*. Agents are situated in a dynamic and complex environment, in which they react promptly.
- *Pro-activeness*. It is not only at an explicit command of the user, that the agent acts: it is aware of its goals and the needs and preferences of its user, and acts accordingly.
- *Social*. Agents are capable of communicating with other agents and with users, and are able to co-operate or compete with each other.

On top of this, agents may be *adaptive* in the sense that they learn about the environment and the user; a *rational* agent is one that acts in its own interest, and *believable* agents are endowed by intelligent and even emotional attitudes

to respond to the user's rich variety of human behavior in an acceptable way (this is an issue in for instance personal assistants on a desktop, but also in the entertainment industry). Finally, *mobile agents* are programs that can migrate from host to host in a network, at time and to places of their own choosing (giving improvement in latency and bandwidth of client-server applications and reducing vulnerability to network disconnection).

Since about 1990, agents have been an important topic in artificial intelligence and computer science, and not only theoretically: more and more systems are implemented with an 'agent-perspective'. In such systems, the notion of agent must be taken in a broad sense: human users are modeled as agents, pieces of software are subdivided in (sub)agents, in applications like traffic and transport agents can be human drivers, a car, a dashboard, but also a traffic light or a junction. Agent-technology has assumed enormous proportions in applications on the Web and in telecom applications. We conclude here with a citation from a leading researcher of British Telecom, taken from [Hermans, 1997]: "Agents are here to stay, not least because of their diversity, their wide range of applicability and the broad spectrum of companies investing in them. As we move further and further into the information age, any information-based organization which does not invest in agent technology may be committing commercial hara-kiri", Hyacinth S. Nwana said.

Information agents

When trying to use the hypertext-oriented information service of the Web, a typical user faces mundane, repetitive tasks such as browsing, filtering and searching for relevant information. Even more data is preserved in connected legacy databases: data that is often volatile, redundant, unstructured and stored in many formats, such as text files, software applications, video and other multimedia systems. The impact of the fast growing information overload makes the tasks of the user, like determining and finding the information sources, dealing with different levels of abstraction and combining partially relevant information, a time-consuming activity.

Information agent technology is an important branch of agent technology, combining techniques and approaches from such divers fields as artificial intelligence, advanced databases, knowledge base systems, distributed information systems, information retrieval and human computer interaction. [Klusch, 2001] defines an information agent as "... an autonomous, computational software entity (an intelligent agent) that has access to one or multiple, heterogeneous and geographically distributed information sources, and which pro-actively acquires, mediates, and maintains relevant information on behalf of users or other agents preferably just-in-time." We can distinguish a number of skills that

are needed to fulfill these goals, and often they are taken care of by a separate agent. Following [Klusch, 2001], we thus distinguish the following skills.

Improving the search for information

One of the first available ways to search for information on the Internet is offered by so-called *search engines*. By means of programs that roam the Internet (with flashy names like *spider*, *worm* or *search bot*) meta-information is being gathered about everything that is available on it. The collected information, characterized by a number of keywords (references) and perhaps some supplementary information, is then put into a large database. Anyone who is searching for some kind of information on the Internet can then try to localize relevant information by giving one or more query terms (keywords) to such a search engine.

However, given that the Internet has become so overwhelmingly large (the estimation for 1999: 200 million users and more than 56 million hosts (for Internet statistics, see [Global Reach, 1999] or [ISC, 1999]. Some novel search engines try to overcome this problem by adding smart features: thus, one can obtain a *meta-search machine* (for instance, *meta-crawler*) that does not search the complete Web, but puts different search engines to work to execute a user's query concurrently. Moreover, some machines offer fancy and efficient ways to present their results to the user, for instance, in a 'five dimensional screen'. It will be clear that such solutions are rather ad hoc and only shift the problem, without solving it.

[Hermans, 1997] mentions the following features which search engines offer, and the improvements that intelligent software agents might offer here:

1 Using a search engine, users must be able to formulate the right set of keywords; using too little, too many or the wrong keywords will cause irrelevant information, or will not retrieve relevant information. Agents, on the other hand, are capable of searching information more intelligently, for instance because tools (such as thesaurus) enable them to search on related terms as well, or even on concepts. On the basis of user information, agents may even fine-tune or correct the user's queries.
2 The way search engines work is very inefficient, causes a lot of data traffic, and it does not account properly for the dynamic nature of the Internet and the information that can be found on it. Agents create their own knowledge base about the available information, recognize when documents move to another location, and are able to cooperate with other agents and use their knowledge.
3 The search for information is often limited to a few sources, such as the Web. Finding information through other services (like in 'Telnet-able' databases)

is left to the user's devices. Agents can relieve their human user of the need to worry about details such as how to operate on specific services.

4 Search engines can't always be reached, servers may be down or too busy. Agents, residing on a user's computer, are always available. Agents can perform several tasks in parallel, day and night. They may even detect and avoid peak hours on the Internet.

5 Search engines are domain-independent. Terms in gathered documents are lifted out of their context and stored as a mere list. Software agents are aware of contexts; deduce it from user-information, or by using other services, such as a thesaurus.

6 The information of the Internet is dynamic, and search engines have problems in adjusting to that. They don't provide the user with updates on one or more topics. Agents, on the other hand, can adjust themselves to preferences and wishes of the user. Moreover, agents can scan the net continuously for updated or even new documents that match the user's interest.

It will be clear that 'conventional software programs' just like humans, will not be capable of comprehending the still growing amount of available information on the Internet. (Note that, although several researchers have argued that an intelligent information agent is comparable with a human librarian [Bundy, 2000; Zick, 2000] there is a notable difference: the latter only searches in indexed databases, whereas software agents can in principle search through complete documents). More intelligent solutions are needed. However, at the moment, although there is a lot of interest in agents, few agent-based programs are available that obey the solutions sketched above. Later in this paragraph we describe a number of implemented systems.

One of the main open problems in the area of information agents is that of *the ontology problem:* how can we guarantee that agents that communicate use some kind of (shared) semantics, rather than just syntax? As [Adolfo, 2000] argues: "In the field of agents, most agents are not built by us, but by somebody else. Thus, our agents will interact mostly with 'unknown and never seen before' agents arriving from all over the planet. These agents will no doubt have diverse goals, and will express their wishes or purpose using different ontologies. Consequently, mechanisms and forms for exchanging information and knowledge among heterogeneous systems are needed." In fact, this problem concerns two issues: agents first of all have to agree on a *communication protocol* and understand each other's so-called *speech acts*. Secondly, agents have to recognize when the *content* of their messages, or between several documents, is (semantically) related. This refers to the problem that documents are written in different (natural) languages, but even with the same language, one often needs semantic information to know that one is talking about one and the same

thing. [Lesser, 1998] calls this the *interpretation problem*, where interpretation is defined as the process of constructing high-level models from low-level data.

In Europe, the FIPA (Foundation for Intelligent Physical Agents [FIPA]) has attempted to produce a standard for agent technologies, including one for a communication language. The language KQML (Knowledge and Query Manipulation Language) is an initiative to achieve the same. [KQML] gives examples of agent 'speaking KQML', a mailing list on the language, and software supporting it. The items 1 and 5 in the list of Hermans, given before, suggest that using a thesaurus may be a good way to solve the second problem; however, there is more to this than finding synonyms for words: agents must for example also understand the relation between measures given in feet or in centimeters. Lenat and Guha have proposed to use the common sense tree of CYC [Lenat, 1989]. Other authors, like [Eijk, 2001] have proposed that translators between different agent systems should evolve dynamically. However, neither of these problems of ontology has been completely solved yet, but we will not address this issue any further here.

At several places in literature [Hermans, 1997; Wong, 2000] it is suggested that one needs to replace the two layers of the Internet (i.e. 'users' and 'suppliers') by a three layer architecture: 'users', 'suppliers' and 'intermediaries'. Each layer can be represented by a number of agents. On the demand or user side agents' tasks would be to find out exactly what users are looking for, what they want, if they have any preferences with regard to the information needed, etc. On the supply side, an agent's tasks would be to make an exact inventory of (the kinds of) services and information that are being offered by its supplier, to keep track of newly added information, etc. Finally, intermediary agents mediate between agents (of the other two layers), i.e. act as (information) intermediaries between (human or electronic) users and suppliers.

An advantage of this approach is that each user can choose his own 'user layer', being either a novice user or an expert, and the same applies for suppliers, which can choose their own tailor-made supply agents. On top of that, [Hermans, 1997] distinguishes the following functions of the middle layer. Firstly, it dynamically matches user demand and provider's supply in the best possible way. Suppliers and users (i.e. their agents) can continuously issue and retract information needs and capabilities. Secondly, it unifies and processes suppliers' responses to queries to produce an appropriate result. The content of user requests and supplier 'advertisements' may not align perfectly. So, satisfying a user's request may involve aggregating, joining or abstracting the information to produce an appropriate result. Thirdly, it actively notifies users of information changes.

Improving the supply of information

Information activity is composed of both information resources and needs. Where the previous subsection focused on more complex, independent client information processing tasks, there have also been researchers who proposed making resources more sophisticated and interoperable. One such proposal comprises XML[2], in which designers of web-pages add tags with semantic information to the items in a page, in order to facilitate better performance in queries. In this subsection we discuss a proposal made by Harmsze in her thesis 'A Modular Structure for Scientific Articles in an Electronic Environment' [Harmsze, 2000]. In her introduction, she compares the activity of scanning printed papers and putting them as bitmap files on the Web, with the first automobiles which were shaped like coaches to be horse-drawn. Over time, the coach-like appearance of automobiles was gradually replaced by a form more appropriate to new technologies, when the advantages of an aerodynamic design were taken into account. The question she addresses is: "how to 'streamline' electronic journals to make them more suitable vehicles for the information highway."

She focuses on scientific papers in journals on physics. Her proposal boils down to dramatically changing the design of such a paper. Rather than begin with an 'abstract', followed by an 'introduction', leading via some standard sections to the final 'conclusion' and 'references', she judges a *modular* structure more suitable for papers on the Internet. With such a structure, the stored document is in fact not one paper, but a collection of papers, one for each potential user. She identifies the following modules:

1 a module containing *meta-information* (bibliographic data, classification code, abstract with links to other modules);
2 a module taking care of the *positioning* (embedding of the problem, with links to other documents, links to a description of the project);
3 a module called *methods* (with links to a general description of the research method, describing the set-up of the experiment so that similar modules in other documents can benefit from it);
4 a module *results* (with links from each result to the method module);
5 a module *interpretation* (here, the results are interpreted, with links to methods, results and possibly other documents;
6 and a final module *outcome*, (in which a summary is given and future research is mentioned). She also provides guidelines on how to add and use links in such a document. Although her proposal may not be the ultimate form of a paper for every discipline, it seems worthwhile that researchers as authors re-think the way they organize their publication, given the fact that it will be made electronically available.

2 See CD-rom: ..\papers\XML in 10 points.htm

Some examples

In this section, we discuss a few applications and sources on the Internet that may become relevant for researchers. For a more detailed overview, the reader is referred to [Etzioni, 1995]. We already mentioned (meta-)search engines as a tool, and special sites for specific scientific topics. Researchers and developers are now focusing on rather basic agent applications, to prove the technology valid. Examples of such agent applications are: agents which partially or fully handle someone's e-mail; agents which filter and or search through news articles looking for information that may be interesting for a user; and agents that make arrangements for gatherings such as a meeting.

A simple, specific-purpose but powerful example of an existing implementation is MIA, the Molecular Information Agent ([MIA]). It is a web interface to an extensible, object-oriented WWW cross-reference search and retrieval system. According to MIA, its goal is to allow users to search the Internet easily to find all current information for a molecule of interest. It has built-in heuristics that direct it to appropriate resources, and its text parsers allow it to scan database entries for links to other sources. At the present time, the MIA supports search by molecule identification number, gene symbol, sequence, and key word. Given a specific entry, it consults a list of templates that predefines queries to data resources that use the entry as an index. Moreover, the results of a database query are parsed and additional keywords are identified.

Gossip [Gossip] is an application developed by Tryllian, in which a user can send his agent with his query on the Internet. It is an example of a mobile agent application. It is an information retrieval and community building tool that finds web sites, pictures and texts by exchanging information with other agents and consulting on-line databases. Gossip's graphical user interface shows the agents moving around autonomously and communicating directly with users. The system continues to work for the user while he is doing other things, even after his computer has been turned off. Gossip is made up of four basic components: 1. the playground, which is Gossip agents' home base. The playground has a doorway to the Internet: the user can modify it to his connection parameters; 2. the agents that will ask the user what to do and when to be back at the playground; 3. the user's profile, which is used by the agents to exchange information with other users with similar interests and profiles; 4. the user's backpacks, which are information containers filled with the instructions and the results they bring back. Every time an agent is sent out with a backpack, it fills it up with more info on that subject and distributes the user's contribution to the Gossip community.

When an agent is sent out, it travels to a Gossip Meeting Point via the Internet. The agent first checks in at the Front Desk, where it tells the Directory Agent

what it has in its backpack and what it's looking for. The Directory agent then points the agent in the direction of other agents with similar backpacks. These agents then find each other and exchange backpack contents based on keywords and profiles. The agent can also browse through a Meeting Point database created from community contributions. It can even submit its info to various search engines from the Meeting Point, if it does not find what it needs to complete its task from within the Meeting Point. It will continue to search until the time specified for its return has elapsed or its backpack is completely filled.

Finally we discuss two applications of information agents taken from [Hermans, 1997]: Softbot and Info Agent. Softbot (see also [Softbot]) assumes that the documents it visits are well-structured, such as stock quote servers or library databases. Because of this, Softbot need not rely on natural language or information retrieval techniques to 'understand' the information provided by a service. Instead, the Softbot relies on a model of the service for the precise semantics associated with information provided by the service. As a result, the Softbot can answer focused queries with relatively high reliability; the chances of finding relevant information are high and the amount of non-relevant information ('noise') is (relatively) low.

The key idea behind the Softbot is reflected in its name, which is derived from software robot. Its tools consist of UNIX commands such as ftp, print, and mail. Commands like list files and Internet services such as Finger and Netfind are used as sensors to find information. The Internet Softbot is a prototype implementation of a high-level assistant. In contrast to systems for assisted browsing or information retrieval, the Softbot can accept high-level user goals and dynamically synthesize the appropriate sequence of Internet commands to satisfy those goals. The Softbot executes the sequence, gathering information to aid future decisions, recovering from errors, and retrying commands if necessary.

The goal-orientedness of the Softbot is only useful if users find specifying requests to it easier than carrying out activities themselves. The agent has three properties which should make goal specification convenient for users: 1. it understands an expressive goal language and can deal with a goal like "get all of researcher Joe's technical reports that are not already stored locally"; 2. it has a convenient syntax and interface for formulating requests. The Softbot supplies a forms-based graphical user interface and automatically translates forms into the logical goal language; 3. it comes with a mixed-initiative refinement dialogue. The Softbot possesses many, but not all of the agent characteristics of agents as described in the subsection Agents. It is autonomous, goal-oriented, flexible and pro-active. At this moment work is being done to extend the Softbot's collaborative, communicative, adaptive and personality-characteristics.

Info Agent (see [Info Agent]) is a system supporting users in retrieving data in heterogeneous archives and repositories. One single agent, the Info Agent, is the interface between the system and the user. This agent uses a so-called Interface Agent for handling the communication with the user. The latter agent is like a personal assistant who is responsible for handling user needs, and for the connection of the user with the agent(s) that will help him solve his problem. As a result of the distributed and agent-based architecture of the system the whole structure of it can be easily changed or updated by adjusting the Interface Agent only. The Interface Agent is able to reason about the user's requests and to understand what type of need he is expressing: it singles out which of the two other agents in the system is able to solve the current problem and sends to it its interpretation of the query.
These other two agents are the Internal Services Agent and the External Retrieval Agent. The Internal Services Agent knows the structure of the archives available in a given organization: it is in charge of retrieving scientific and administrative data, performing some class of actions (such as finding available printers) and supporting the user in compiling internal forms. The External Retrieval Agent is in charge of retrieving documents on the network. It can work in two modalities: retrieval (or query) mode and surfing mode. In the first case, it searches for a specific document following a query asked by the user: this service is activated by a direct user request. In the second case, the agent navigates the network searching for documents that, in its opinion, could interest the user. The search is driven by a user's profile maintained by the Interface Agent. Hermans observes that the Info Agent resembles, in a number of ways, the Softbot. One of the differences between these two agents is that the Info Agent focuses mainly on the user, whereas the Softbot focuses mainly on the requests of the user. Another difference is that the Info Agent searches in both structured as well as unstructured information (documents), whereas the Softbot 'limits' itself to structured information only.

Conclusion

This paper describes some technologies that aim to assist the user, and in particular the scientist, in finding and managing the information that becomes (electronically) available. In doing so, we zoomed in to the technology offered by so-called *agents*, which, allegedly, show at least a minimum of intelligence in doing their task. We saw that the agent community makes fairly firm claims about their stance, and the literature on agents is indeed immense. On the other hand, existing information agents are still scarce, but both companies and researchers expect that they offer a promising way to follow. In some sense, this is an exciting time, and it is too early to judge developments. The interested reader is encouraged to consult some of the valuable sites and sources mentioned in the bibliography.

References

- Adolfo, G.A., J.M. Olivares Ceja, A. Ma. del Carmen Dominguez. (2000). Agents that Interact Using Mixed Ontologies. Manuscript. Unpublished
- Bundy, A. (2000). Drowning in Information, Starved for Knowledge: Information Literacy, not Technology, is the Issue. Paper presented at Books and Bytes: Technologies for the Hybrid Library 10th VALA Conference. Melbourne 16-18 February. http://www.library.unisa.edu.au/papers/drowning.htm
- Berners-Lee, T. (1999). World Wide Web seminar. http://www.w3.org/Talks/general.html
- Dennet. D.C. (1987). The Intentional Stance. The MIT Press, Cambridge, MA
- Eijk, R. van, F. de Boer, W. van der Hoek, J.-J.Ch. Meyer. (2001). On Dynamically Generated Ontology Translators in Agent Communication. Accepted for the Journal of Intelligent Systems
- Etzioni, O., D. S. Weld (1995). Intelligent Agents on the Internet - Fact, Fiction, and Forecast. IEEE Expert **4**:44-49
- Garvey, W.E. (1979). Communication: the Essence of Science - Facilitating Information Exchange among Librarians, Scientists, Engineers and Students. Pergamon Press, Oxford
- FIPA. Foundation for Intelligent Physical Agents. At http://www.fipa.org
- Global Reach. (1999). Global Internet Statistics (By Language). See also the electronic version: http://glreach.com/globstats
- Gossip. http://www.tryllian.com
- Harmsze, F.A.P. (2000). A Modular Structure for Scientific Articles in an Electronic Environment. Ph.D. Thesis. University of Amsterdam
- Hermans, B. (1997). Intelligent Software on the Internet: An Inventory of Currently Offered Functionality in the Information Society and a Prediction of (Near) Future Developments. First Monday **2**:3. http://firstmonday.org/issues/issue2_3/ch_123/
- Hofstadter, D. (1981). Metamagical Themas: A Coffeehouse Conversation on the Turing Test to Determine if a Machine Can Think. Scientific American. May. pp15-36
- IEEE. (1999). INSPEC, the Quality Database for Physics, Electronics and Computing. http://www.iee.org.uk./publish/inspec/
- Infoagent. http://www.cilea.it/GARR-NIR/nir-it-95/atti/giannini/giannini/giannini-nir-95.html
- ISC. (1999). Internet Software Consortium, Internet Domain Survey. At http://www.isc.org.ds/
- Klusch, M. (2001). Information Agent Technology for the Internet: A Survey. Journal of Data and Knowledge Enigineering. Forthcoming. http://www.dfki.de/~klusch/papers/iat-dke-2000.zip
- KQML. http://www.cs.umbc.edu/kqml/

- Leiner, B.M., V.G. Cerf, D.D. Clark, R.E. Kahn, L. Kleinrock, D.C. Lynch, J. Postel, L.G. Robers, S. Wolff. (1998). A Brief History of the Internet. http://www.isoc.org./internet-history/brief.html
- Lenat, D.B., R.V. Guha. (1989). Building Large Knowledge-Based Systems. Addison Wesley
- Lesser, V., B. Horling, F. Klassner, A. Raja, T. Wagner, S. Zhang. (1998). BIG: A Resource-Bounded Information Gathering Agent. http://firstmonday.org/
- Meadows, A.J. (1974). Communication in Science. Butterworths, London
- Meadows, A.J. (1998). Communicating Research. Academic Press, San Diego
- MIA. http://mia.sdsc.edu/
- Naisbett, J., P. Aburdene. (1980). Megatrends. William Morrow New York
- Neumann, J. Von, O. Morgenstern. (1944). Theory of Games and Economic Behaviour. Princeton University Press
- Softbot. http://www.cs.washington.edu/research/projects/softbots/www/softbots.html
- Stich, G.P. (1983). From Folk Psychology to Cognitive Science. The MIT Press, Cambridge, MA
- Turnbull, H.W., J.F. Scott, A.R. Hall. (eds). (1956). The Correspondence of Isaac Newton. Cambridge University Press, Volume I:1661–1675. http://www.newtonia.freeserve.co.uk/E/Giants.html
- Wong, H.C., K. Sycara. (2000). A Taxonomy of Middle-Agents for the Internet. http://firstmonday.org/
- Wooldridge, M., N.R. Jennings. (1995). Intelligent Agents: Theory and Practice. The Knowledge Engineering Review **10** (2):115-152
- Wooldridge, M., A. Rao. (eds.). (1999). Foundations of Rational Agency. Kluwer Academic Publishers, Boston, MA
- Zick, L. (2000). The Work of Information Mediators: A Comparison of Librarians and Intelligent Software Agents. http://firstmonday.org. issues/issue5_5/zick/index.html

2.2.3 Science mapping from publications

An example in mathematics & computer science [1]

Ed Noyons and Ton van Raan [2]

Introduction

In this Section we discuss the creation of 'maps of science' with help of advanced bibliometric methods. This 'bibliometric cartography' can be seen as a specific type of data mining, applied to large amounts of scientific publications. As an example we describe the mapping of the field mathematics & computer science (MCS). The mapping is based on 'co-word analysis' [Callon, 1983; Noyons, 1998] and applied to CompuMath, the special Citation Index of ISI [3] for computer science and mathematics. The number of publications covered by this database is about 50,000 per year.

This article addresses the main lines of the methodology. We will illustrate the results with a project carried out for the Swiss Science and Technology Council. The aim of the project was to map the field and to assess the 'position' of the Swiss MCS research for the period 1995-1998. Current research is going on to update the mapping for the years 1999-2001.

Each year about a million scientific articles are published. For just one research field, such as MCS, the amount of papers is already about fifty thousand per year. How is it possible to keep track of all these developments? Are there cognitive structures and patterns 'hidden' in this mass of published knowledge, at a 'meta-level'?

Suppose each research field can be characterized by a list of most important, say 200, keywords or, in most cases, phrases i.e. keyword-combinations ('concepts'). For MCS research such a list will cover words like differential equation, optimization, chaos, fuzzy set, parallel computer, Monte Carlo simulation, and so on. Each publication can be characterized by a subset from the total list of keywords. It is, as it were, a DNA fingerprint of a publication. For all publications, keyword-lists are compared pair-wise. In other words, these many thousand publications constitute a gigantic network in which all publications are linked together by one of more common keywords. The more keywords two publications have in common, the more these publications are related (keyword-similarity) and thus belong to the same research area or research specialty. In the biological metaphor: the more DNA two objects have in common, the more they are related. Above a certain similarity threshold, they will belong to a specific species.

Mathematical techniques are used to unravel these publication networks, by word-similarity measurements, clustering of related publications, and finally

1 This article is based on a report for the Centre of Science and Technology Studies (CEST) attached to the Swiss Science and Technology Council, Bern

2 Dr E.C.M. Noyons, noyons@cwts.leidenuniv.nl and Prof Dr A.F.J. van Raan, vanraan@cwts.leidenuniv.nl, Centre for Science and Technology Studies (CWTS), Leiden University, The Netherlands, www.cwts.leidenuniv.nl

3 The Institute for Scientific Information in Philadelphia, the publisher of the Science Citation Index and all other related citation indexes.

mapping the ensemble of these clusters in a two-dimensional space. This procedure visualizes an underlying structure. The fascinating point is that these structures can be regarded as the cognitive, or intellectual structure of the scientific field. Clusters can be identified as subfields and research themes.
As discussed above, the procedure is entirely based on the total of relations between all publications. Thus, the structures that are discovered are not the result of any pre-arranged classification system. The structures emerge solely from the internal relations of the whole universe of publications together.
In other words, what is made visible by our mathematical methods, is the self-organized structure of science. A detailed discussion of science maps based on co-word analysis is given in a recent publication [Noyons, 1999].

Methodology

From the above discussion it is clear that keywords from publications play a central role in the methodology. Only noun phrases (NPs) can become field 'keywords'. In order to identify noun phrases in English texts, we use a computer-linguistics based 'noun phrase extractor' (the 'parser'). The identified NPs are divided into two groups: the single word NPs (SWNP) and the multiword NPs (MWNP). From the list of MWNPs, those with too general a meaning are removed.
The selection of field-specific keywords from the list of remaining MWNPs is presently established on the basis of their frequency distribution, and (if possible) the input of field experts. For each MWNP, we count the number of occurrences in titles and abstracts within the field under study, as well as the number of occurrences in titles in science as a whole (i.e. all publications (about a million!) covered by all ISI citation indexes). On the basis of these results the specificity of the NP within the field and its 'centrality' within the field is determined (see [Noyons, 1999] for a detailed discussion).
By using an 'on-line' feedback form, experts are enabled to remove preliminary selected keywords or to add preliminary excluded NPs from the two lists. In order to identify clusters (subfields) within a field, we first construct a matrix composed by co-occurrences of the selected keywords (about 900) in the set of publications for a specific period of time (we start with the most recent period, in the example: 1997-1998). We normalize this 'raw co-occurrence' matrix in such a way that the similarity of keywords is no longer based on the pair-wise co-occurrences, but on the co-occurrence 'profiles' of the two keywords in relation to all other keywords.
This similarity matrix is input for a cluster analysis. In most cases, we use a standard hierarchical cluster algorithm including statistical criteria to find an optimal number of clusters. The identified clusters of keywords represent subfields. These subfields are labeled with a name by the four most frequent keywords in a cluster.

To construct a map of the field, the subfields are positioned in a two-dimensional space. Each subfield represents a subset of publications on the basis of the above discussed keyword similarity profiles. If any of the keywords is in a publication, this publication will be attached to the relevant subfield. Thus, publications may be attached to more than one subfield. The overlap between subfields in terms of joint publications is used to calculate a further co-occurrence matrix based on subfield publication similarity.
The subfields are positioned in two-dimensional space by multidimensional scaling. Thus, subfields with a high similarity are positioned in each other's vicinity, and subfields with low similarity are distant from each other. The size of a subfield (represented by the surface of a circle) indicates the share of publications in relation to the small number in the field as a whole. Particularly strong relations between two individual subfields are indicated by a connecting line. As discussed above, we begin our mapping procedure with the data for a recent time period (here 1997-1998).

The maps can be published on the CWTS web site [4]. Through this browser based interactive interface the maps can be explored and validated. Information 'behind' the map is provided in the same way (actors, and their output and impact indicators).

The map created by our co-word based methodology does not cover 100% of the MCS publications in Compumath. In other mapping projects, for instance neuroscience, we reach a coverage of 80% or more. In this field, we cover only 60% of the publications in 1997-1998. Most probably this relatively low coverage is related to the communication characteristics of the field. Mathematics abstracts contain a lot of 'non-language' expressions such as symbols and formulas. Therefore, less keywords are available for the co-word analysis.

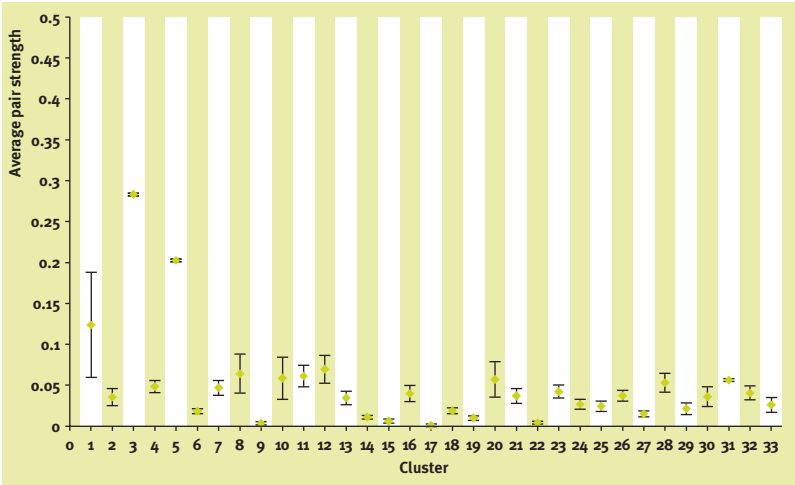


Figure 1
Mathematics Computer Science internal cluster coherence (1997-1998).

4 http://www.cwts.leidenuniv.nl

Analysis mathematics and computer science

The clusters resulting from the mapping procedure have been tested for internal coherence. We calculated the average linkage between all keyword (NP)-pairs within a cluster, and the standard deviation. This internal coherence measure indicates the robustness of the identified subfield. The results are given in Figure 1.

Legend of Figures 1 to 6

subfields	
1	ATM/ performance evaluation
2	Linear model/Bayesian approach/ Monte Carlo method
3	Case based reasoning
4	Fuzzy set/membership function
5	Large number/strong law
6	Simulation study/ maximum likelihood/asymptotic distribution/ distribution function
7	Robotics
8	Scheduling problem/single machine/traveling salesman problem/ job shop
9	Artificial neural network/computer simulation/ computational complexity/ expert system
10	Stability/traveling wave
11	Parallel computer/parallel algorithm
12	Necessary condition/optimal control
13	Optimization
14	Boundary condition/numerical simulation/numerical experiment/partial differential equation
15	Dynamical system/chaos/time series/initial condition
16	Internet/ WWW/website
17	N 1/Finite Group/Monte Carlo simulation
18	Discrete time/continuous time/frequency domain/nonlinear systems
19	Necessary and sufficient condition/Banach Space/Hilbert space/ l 2
20	Vertex/regular graph/chordal graph/distance regular graph
21	Genetic algorithm/objective function/simulated annealing/tabu search
22	Asymptotic behavior/boundary value problem/approximate solution/exact solution
23	Finite element
24	Real time/petri net/ formal method
25	Differential equation/2nd order/runge kutta method/first order
26	Initial data/cauchy problem/global existence/initial boundary value problem
27	Polynomial time/linear time/approximation algorithm/bipartite graph
28	R n/ positive solution/bounded domain/semilinear elliptic equations
29	Complexity/lower bound/upper bound/ branch and bound algorithm
30	Classification/feature extraction/discriminant analysis/texture classification
31	Robustness/disturbance rejection
32	Numerical method/finite difference method
33	Sufficient condition/asymptotic stability/lyapunov function/global stability

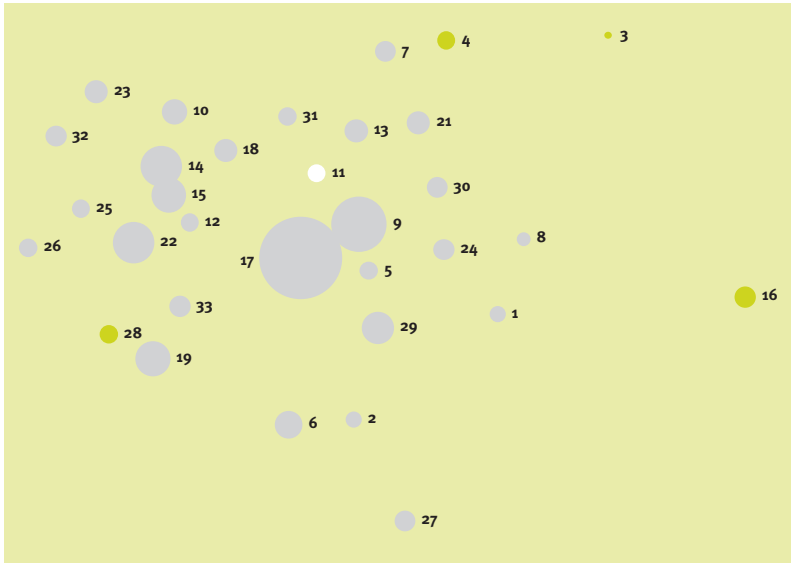


Figure 2
Map of Mathematics & Computer Science (1997-1998). Two-dimensional representation based on the similarities between identified clusters of keywords (subfields). For the list of subfields with corresponding number we refer to the legend of Figure 1. The size of the subfields represents the number of publications in a specific subfield. The color indicates a significant change of publication activity. Green: increase of activity; White: decrease of activity. The badness-of-fit criterion is 0.22, the distance correlation is 0.88.

In Figure 2 we present the structure of MCS by a 2D map of the structural relations between subfields.

As discussed in the methodology section, this map is a two-dimensional representation of a structure resulting from the cognitive, i.e. keyword/concept-based relations of subfields, measured by the co-occurrence of these concepts. These subfields are defined by clusters of related keyword/concepts. By mapping the structural relations between subfields of successive time periods, we create a 'movie' of the evolution of the field within that period. In this movie (see for examples our CWTS web site), we visualize the evolution of individual subfields (growth, in terms of publications), and of their relations with each other (positioning). The information 'behind' the map, can be explored through an interactive browser based interface. Selection of a specific information 'option', enables the user to retrieve data by clicking the relevant subfield circles. The following options are available for each subfield: the most frequently publishing authors, organizations, and countries, as well as the most frequently used journals, the most highly cited publications, authors and organizations. In addition, information is provided to evaluate and validate the structure itself on the basis of word- and citation-linkages between subfields.

In the next steps, we can investigate:
- the relative share of a particular country or institution within the science field.
- The development of activity of this country.
- The impact of the publications of a country or institution compared to the world average within the field.

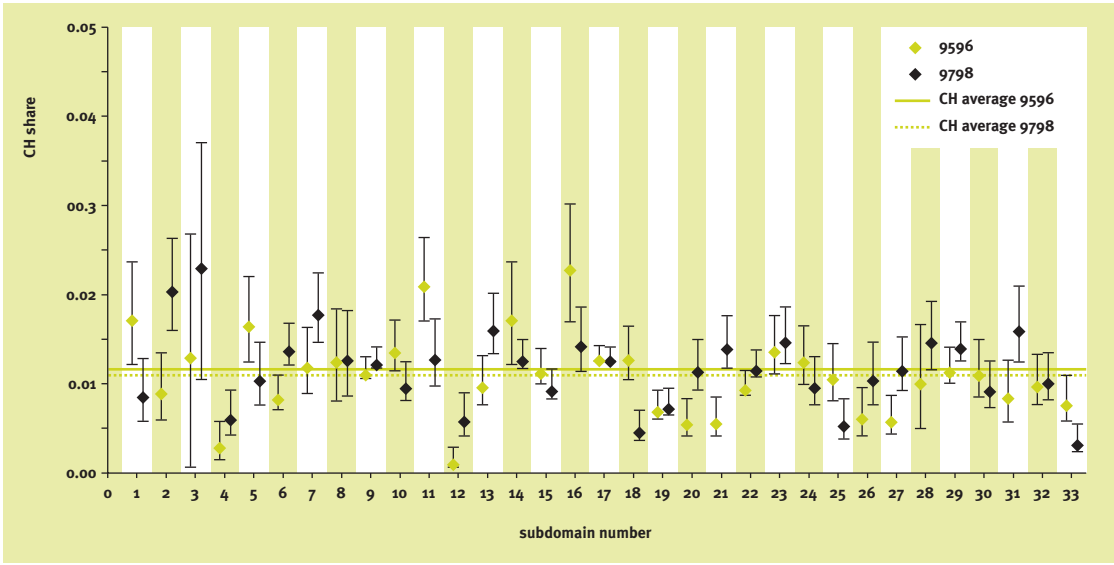


Figure 3
Swiss share in Mathematics & Computer Science subfields (1995-1996 and 1997-1998). For subfields see Figure 1.

These steps are illustrated below with the results of a project carried out for the Swiss Science and Technology Council.

Relative share

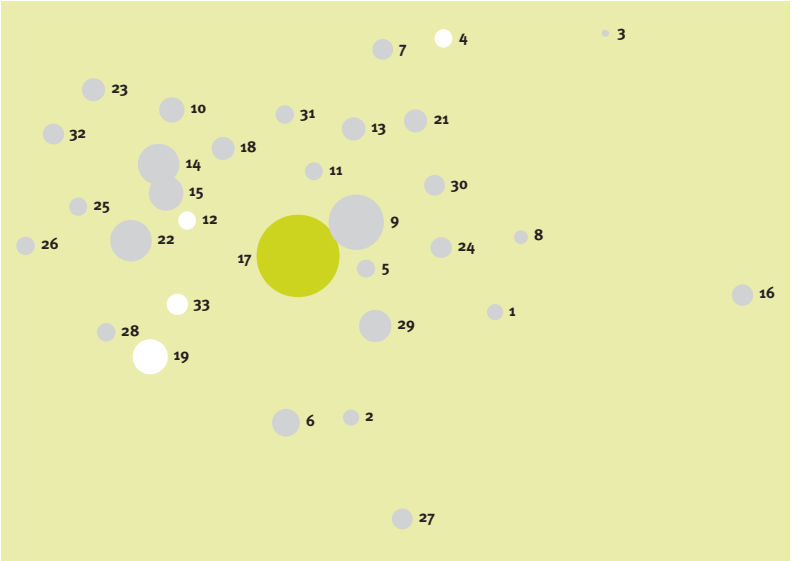
To get an overview of the Swiss activity distribution over the field of MCS, we calculated the share of publications with at least one Swiss address per subfield. This share is a percentage of the total number of publications in a subfield. We determined the Swiss share for the two used periods of time (1995-1996 and 1997-1998) to provide an indication of a trend in the Swiss activity. The error bars added to the data-points indicate the significance of the identified trend. The results are presented in Figures 3 and 4.

We find that the average share of Switzerland in MCS is somewhat more than 1% of the world's total output. In one subfield (17: Finite Group, Monte Carlo simulation), the share of Swiss activity is above this average in the whole period, i.e. both in 1995-1996 and in 1997-1998. This central subfield, in which much general research is covered, Switzerland shows an interest that is significantly above its average in the whole field. In subfields 4, 12, 19 and 33, the Swiss activity share is below its own field average in the whole period.

Development within the subfield

There are two subfields in which the Swiss share decreases significantly in the studied period (18 and 33). There are three subfields (1, 11, 25) in which the decrease remains only just within the calculated error bars. In the case of 11, we are dealing with a subfield with a significantly decreasing world wide interest. In all five subfields with a Swiss share decrease, the absolute Swiss output

Figure 4
Map of Mathematics & Computer Science research with indication of the Swiss share in subfields (1995-1996 and 1997-1998) Colors indicate a Swiss share significantly above (Green) or below (White) its own field average throughout the whole period. For subfields see Figure 1.



decreases as well. There is one subfield (16), in which the Swiss share decreases (though just within the error bars) but in which the absolute number of Swiss output *increases*. In this particular subfield, the world activity increase exceeds the Swiss increase. The conclusion that Switzerland does not keep up with the pace world wide (primarily US publications) is too simple. In the earlier period (1995-1996) Switzerland already showed a relatively high share (more than 2%). In the later period (1997-1998) its activity is lowered to a more average Swiss level (around 1%). The fact that the world-wide activity was increased in 1997-1998, could therefore also be interpreted as a good foresight of the Swiss researchers in this area. We stress, however, to be careful with conclusions based on relatively low absolute numbers.

Furthermore, there are seven subfields in which the Swiss share increases significantly (2, 6, 12, 13, 20, 21, 27). In all these cases the world-wide interest increases as well. Four of these subfields are located at the 'lower' side of the map. Apparently, Swiss MCS research has directed its focus to this area of the field. Swiss activity is also increased in the area above the center. We already mentioned 13 and 20, but also in 3, 4, 7 and 31 an increase of Swiss activity is noted, although not exceeding the error bars. In Figure 5, we summarize the increasing/decreasing share of Switzerland in 1997-1998 in relation to 1995-1996.

Impact

Finally, we indicated in the map those subfields in which Swiss research in MCS reaches an impact that is significantly above or below the world average in 1995-1996. The world average is determined by the average impact of a publication per subfield. Figure 6 shows that the strength of Swiss MCS research is in the core area of the field, i.e. in and around 17 (1, 7, 9, 10, 12, 18, 19, 28, 29, and 33).

Figure 5
Map of Mathematics & Computer Science research with the development of Swiss share in subfields (1995-1996 and 1997-1998).
Color legend:
Green: increase outside error bars.
Dark grey: share increase in 1997-1998 over 20% of 1995-1996.
Black: share decrease in 1997-1998 over 20% of 1995-1996.
White: decrease outside error bars.
For subfields see Figure 1.

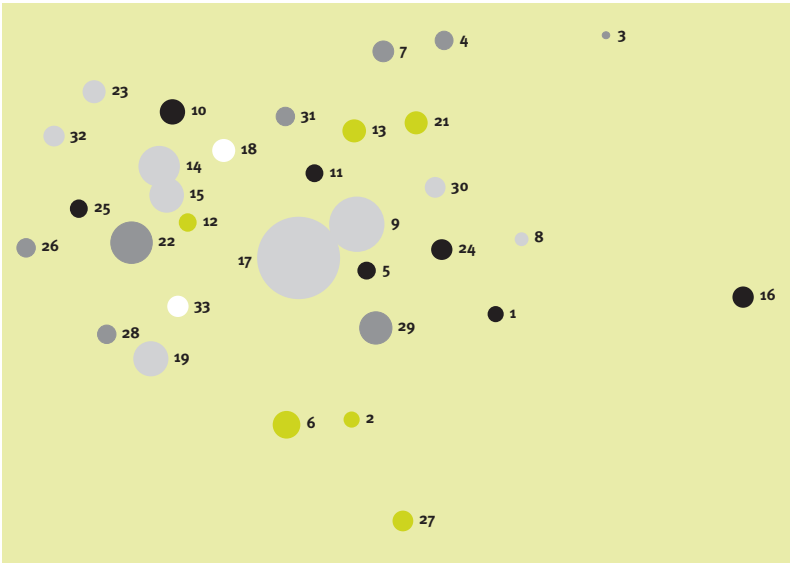
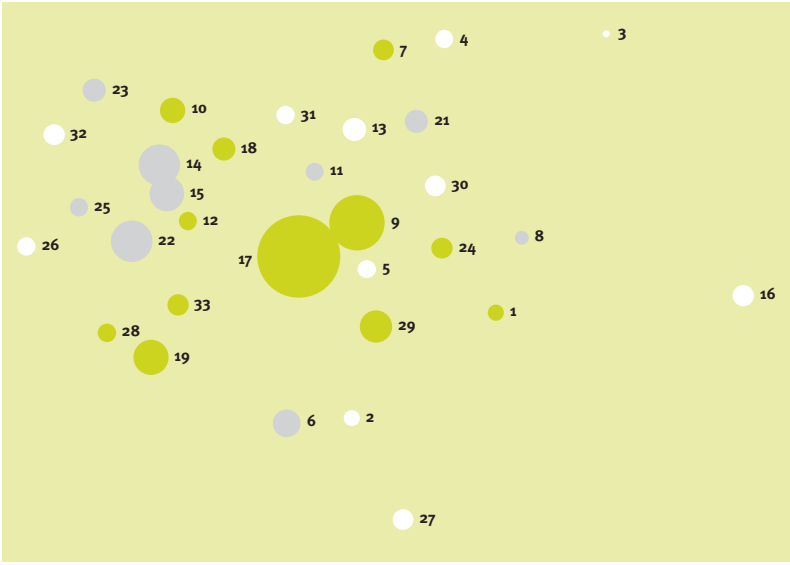


Figure 6
Map of Mathematics & Computer Science research with the Swiss impact as compared to the world average in subfields (1995-1996).
Color legend:
Green: Swiss impact higher than 1.2 times world average.
White: Swiss impact lower than 0.8 times world average.
For subfields see Figure 1.



In general, the Swiss impact in the 'periphery' is lower. In subfields 2, 3 and 4 it is even zero.

Bibliometric analysis of the underlying publication data allows us to identify high-impact institutions. Such specific results can be looked up in more detail via the CWTS map interface.

A detailed discussion of our bibliometric method to measure impact on the basis of citation analysis is given in recent publications [Noyons, 1999; Raan, 1996, Raan, 1999; Raan, 2001].

Conclusion and future vision

With bibliometric mapping we are able to depict the cognitive structure of scientific fields. These cognitive, semantics-based structures act as a 'basic landscape' in visualizing the mutual relations and linkages between subfields and themes within science fields, as well as the *interdisciplinary* relations with other fields. By introducing a time-dimension in the analysis (time-series of maps) we are able to identify *newly emerging themes* ('dynamics of the field'). Thus bibliometric mapping helps to answer crucial questions such as: how does an R&D field look in terms of its cognitive, intellectual structure? How is the field related to its direct 'scientific environment'. Is it possible to explore this 'scientific environment' from the perspective of socio-economic problems? Who and where are the important actors?

Moreover, given the generic character of the methodology, our approach can be extended, if appropriate, immediately to any other data system of (electronically available) documents (e.g. patents, reports, proposals) and databases covering the most recent important international conferences, as well as databases compiled from appropriate sources available via Internet and electronic publishing.

References

- Callon, M., J.-P. Courtial, W.A. Turner, S. Bauin. (1983). From Translations to Problematic Networks: an Introduction to Co-Word Analysis. Social Science Information **22**:191-235
- Noyons, E.C.M., A.F.J. van Raan. (1998). Monitoring Scientific Developments from a Dynamic Perspective: Self-Organized Structuring to Map Neural Network Research. Journal of the American Society for Information Science (JASIS) **49**:68-81
- Noyons, E.C.M. (1999). Bibliometric Mapping as a Science Policy and Research Management Tool. Thesis Leiden University. DSWO Press, Leiden
- Noyons, E.C.M., M. Luwel, H.F. Moed. (1999). Combining Mapping and Citation Analysis for Evaluative Bibliometric Purposes. Journal of the American Society for Information Science (JASIS) **50**:115-131
- Raan, A.F.J. (1996). Advanced Bibliometric Methods as Quantitative Core of Peer Review Based Evaluation and Foresight Exercises. Scientometrics **36**:397-420
- Raan, A.F.J. (1999). Scientific Excellence of Research Programs as Pivot of Decision-making. The IPTS Report 40:30-37. Institute for Perspective Technological Studies, Joint Research Institute, European Commission, Seville
- Raan, A.F.J., Th. N. van Leeuwen. (2001). Assessment of the Scientific Basis of Interdisciplinary, Applied Research. Application of Bibliometric Methods in Nutrition and Food Research. Research Policy, to be published

2.2.4 Mining for scientific hypotheses

Jan Rauch[1]

Introduction

The goal of this Section is to outline a selection of methods for generating scientific hypotheses through data mining. The generation of hypotheses can be seen as a step in scientific methodology. Usually, large quantities of data and a general problem like *What are the possible causes of a given phenomenon?* are on the input side. The result will consist of several hypotheses that are supported by the given data. The core of the step is automatic formulation and testing of a very large amount of potentially interesting hypotheses.

First, we will outline the main principles of the proposed method. Secondly, we discuss the mathematical approach, and some special tools suitable for mining scientific hypotheses. The last paragraph contains some remarks to further possibilities for applying the introduced results.

In a tutorial included on the CD-rom we show two examples of hypothesis mining. This includes hypothesis mining based on statistical tests, as discussed In this article. We will provide a 'hands on' guide, so you can experiment with the software included on the disk.

Theory of computerized hypothesis formation

The ultimate question related to the theory of computerized hypothesis formation presented in the monograph [Hájek, 1978] is the question *Can computers formulate and justify scientific hypotheses?* In this book, the logic of discovery is suggested to deal with this question. This logic can be divided into a logic of induction and a logic of suggestion. The logic of induction studies the notion of justification of hypothesis. The logic of suggestion studies methods of suggestion of reasonable hypotheses.

The theory presented in this book is based on the following scheme of inductive inference:

theoretical assumptions, observational statement
theoretical statement

Figure 1
Scheme of inductive inference.

It means that if we accept theoretical assumptions and verify a particular statement concerning observed data, we accept the theoretical statement forming the conclusion. It is very important that suitable statements concerning observational data lead to theoretical conclusions, not to the data themselves. The languages that are used to express observational and theoretical statements can be considered crucial elements in the process. The following questions can be formulated [Hájek, 1978].

1 Doc.RNDr J. Rauch CSc., rauch@vse.cz, Faculty of Informatics and Statistics, University of Economics, Prague, Czech Republic

The logic of induction

1 In what languages does one formulate observational and theoretical statements? (What is the syntax and semantics of these languages? What is their relation to the classical first order predicate calculus?)
2 What are rational inductive inference rules bridging the gap between observational and theoretical sentences? (What does it mean that a theoretical statement is justified?)
3 Are there rational methods for deciding whether a theoretical statement is justified (on the basis of given theoretical assumptions and observational statements?)

The logic of suggestion

4 What are the conditions for a theoretical statement or a set of theoretical statements to be of interest (importance) with respect to the task of scientific cognition?
5 Are there methods for suggesting such a set of statements which is as interesting as possible?

A mathematical logic of discovery has been developed; both observational and theoretical statements are formulas of special logical calculi. The syntax and semantics of these logical calculi are defined. The formulas refer to mathematical structures and it is defined when the formula is true and when it is false in a given structure.

Observational and theoretical statements

Observational calculi are distinct from the theoretical ones. A typical feature of the observational calculus is the possibility of an effective decision if a given formula is true in a given structure. The effectiveness is defined in a mathematical way. A typical feature of the theoretical calculus is that it is related to the system of possible worlds.
Observational and theoretical calculi are related by inductive inference rules. The scheme of inductive inference rules is given in Figure 2. Statistical hypothesis testing can be used to express rationality criteria of the inference rules. Rational methods for deciding whether a theoretical statement is justified (see question 3) are based on the computational effectiveness of the used statistical procedures.

Theoretical statements are in one-to-one correspondences with a number of specific observational statements for many important inductive inference rules. Thus, the search for theoretical statements (hypotheses) can be reduced to the search for observational statements. This leads to a definition of an observational research problem. The observational problem is given by a set of observa-

tional questions. A solution of the observational problem is a representation of all true observational statements.

Observational calculi

Various types of observational calculi are defined in [Hájek, 1978]. The observational predicate calculus is an example. Informally speaking, a formula of the observational predicate calculus consists of two derived Boolean attributes connected by a generalized quantifier. Formulas:

$$age(35–40) \wedge sex(F) \wedge syst(>150) \Rightarrow^{!}_{0.95;0.05;100} diag(D_Z)$$

and

$$age(50–60) \wedge sex(M) \wedge diag(D_Z) \sim_{0.1,50} syst(>160)\ diast(>90)$$

can be understood as examples of formulas of the observational predicate calculus. Properties of observational predicate calculi are also studied in [Rauch, 1997a; Rauch, 1998a]. The relationships among observational calculi, the GUHA method and fuzzy sets were also studied, see e.g. [Hájek, 1998] and [Iványek, 1999].

Hypotheses mining with the GUHA method

In the past 35 years of research in mathematical statistics and mathematical logic a method has been developed to assist in the mining of hypotheses. It has been named GUHA (General Unary Hypothesis Automaton) [Hájek, 1978]. The GUHA method has been implemented several times e.g. [Hájek, 1995; Rauch, 2000]. A recent implementation is in the 4ft-Miner procedure, which is part of the LISP-Miner software package that is used in the tutorial.

The aim of the GUHA method is to give all the interesting relations that follow from the analyzed data, focusing on the given problem. The method is divided in several procedures. It is implemented in a computer program that works according to the general framework of many KDD methods, as shown in Figure 2.

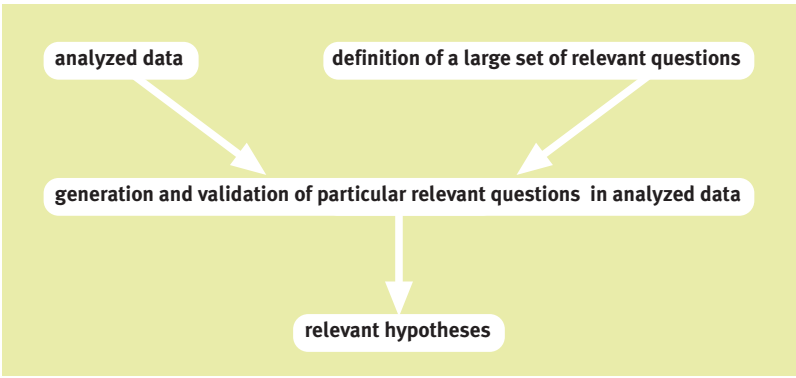


Figure 2
Diagram of the GUHA procedure.

Input data

The analyzed data is handled in the form of a data matrix. This matrix consists of an index value and values for all observed attributes plus two additional Boolean attributes. An example of the analyzed data matrix M concerning patients is given in Table 1.

patient	**observed attributes**						**Boolean attributes**	
	age	sex	diagnosis	systolic blood pressure	diastolic blood pressure	...	φ	ψ
o_1	35	F	D_Z	140	80	...	1	0
o_2	49	M	D_A	180	110	...	0	1
...	...	...	...	...	...	...	...	...
o_n	21	F	D_C	120	90	...	1	0

Table 1
The analyzed data matrix M. The rows of the data matrix correspond to observed objects or situations (patients o_1, ..., o_n). The columns of the data matrix correspond to observed values (e.g. age, sex, diagnosis, systolic blood pressure, diastolic blood pressure, level of sugar). The first patient is a 35 year old woman with diagnosis D_Z, systolic blood pressure 140, diastolic blood pressure 80, etc.

Hypotheses and Boolean attributes

Two Boolean attributes have been added to the data matrix: φ and ψ. Boolean attribute φ is called the antecedent, Boolean attribute ψ is called the succedent. Both are derived from the observed attribute values.

GUHA mines for hypotheses of the form $\varphi \approx \psi$. The symbol $\approx$ is called a generalized quantifier. It is a name of a relation between φ and ψ. The relation $\varphi \approx \psi$ can be true or false for the analyzed data matrix. Some generalized quantifiers correspond to statistical hypothesis tests. If a relation $\varphi \approx \psi$ is true at the analyzed data, then the corresponding hypothesis is supported by the analyzed data.

Example of Boolean attributes

An example of the Boolean attribute derived from the observed attributes is the Boolean attribute age(35–40) $\wedge$ sex(F) $\wedge$ diag(D_Z) (with $\wedge$ meaning 'and'). The basic Boolean attribute age(35–40) is true (indicated by '1') for patients 35–40 years old etc. Thus, the derived Boolean attribute age(35–40) $\wedge$ sex(F) $\wedge$ diag(D_Z) is true for female patients, aged 35–40 years with diagnosis D_Z.

The relation $\varphi \approx \psi$ is verified on the basis of a four-fold table of φ and ψ on the analyzed data matrix (see Table 2).

Table 2
The four-fold table of φ and ψ on M. a is the number of objects satisfying both φ and ψ, b is the number of objects satisfying φ and not satisfying ψ, c is the number of objects not satisfying φ and satisfying ψ, d is the number of objects satisfying neither φ nor ψ. r = a + b is the number of objects satisfying φ, similarly for s, k, and l, n is the number of all objects.

	ψ	$\neg\,\psi$	
φ	a	b	r
$\neg\,\varphi$	c	d	s
	k	l	n

Set of relevant questions

The input of the GUHA procedure consists of the analyzed data matrix and of a definition of a large set of relevant questions (see Figure 2). A relevant question corresponds to a potentially interesting hypothesis $\varphi \approx \psi$. It can be understood as a question *Is the hypothesis* $\varphi \approx \psi$ *supported by the analyzed data?* The set of relevant questions is given by:

- a generalized quantifier $\approx$, e.g. $\Rightarrow^{!}_{p;\alpha;Base}$ (logically implies with a probability p and a significance level α) or $\sim_{\alpha,Base}$ (is not with a confidence level α).
- a set of relevant antecedents (Boolean attribute φ in $\varphi \approx \psi$).
- a set of relevant succedents (Boolean attribute ψ in $\varphi \approx \psi$).

The set of relevant antecedents is given by a set of antecedent attributes and by a minimal and maximal number of attributes in the antecedent. An example of the set of antecedent attributes is the set {age, sex, diag}. A simple definition of a set of basic Boolean attributes is given for each antecedent attribute.

Example of a simple definition of basic Boolean attributes

The set diag(D_A) diag(D_B), ..., diag(D_Z) of basic Boolean attributes is defined by the simple definition 'use particular values'. The set diag(D_A,D_B), diag(D_A,D_C), ..., diag(D_A,D_Z), diag(D_B,D_C), ..., diag(D_Y,D_Z) of basic Boolean attributes is defined by the simple definition 'use pairs of particular values', etc. The basic Boolean attribute diag(D_A,D_Z) is true for patients 01 and 02 (their diagnosis is D_A or D_Z). The basic Boolean attribute diag(D_A,D_Z) is false for patient on (his diagnosis is neither D_A nor D_Z), See the data matrix in Figure 2.

There are a lot of further possibilities to define the set of basic Boolean attributes for each antecedent attribute. Various possibilities of using intervals of values are also included. The relevant antecedent is a conjunction of basic Boolean attributes, e.g. age(10–20) ∧ sex(M) ∧ diag(D_A), ..., age(10-20) ∧ sex(M) ∧ diag(D_Z), age(10–20) ∧ sex(F) ∧ diag(D_A, D_B), ..., etc.

Generating relevant hypotheses

Usually thousands of antecedents are defined. The set of relevant succedents is defined in a similar way. The GUHA procedure automatically generates each particular relevant question $\varphi \approx \psi$ (potentially interesting hypothesis) and tests if the hypothesis $\varphi \approx \psi$ is supported by the analyzed data. The output consists of all interesting hypotheses supported by the analyzed data. When no hypothesis have been generated, it is possible to deduce that there is no interesting hypothesis concerning the given problem that is supported by the analyzed data.

The number of output hypotheses usually ranges from a few to several hundreds. It can be influenced by parameters expressing the intensity of searched

relations. Additional software is available for sorting and filtering of output hypotheses.

The procedure generates and tests hypothesis very fast even when complex statistical tests are used. To achieve this, some special tools are used: analyzed data are represented by suitable strings of bits [Rauch, 1978], and tables of critical frequencies are used to convert a verification of complex statistical tests to tests of a simple inequality [Rauch, 1998b].

Hypotheses testing

GUHA validates hypotheses on the basis of statistical tests (e.g. lower critical implication $\Rightarrow^{!}_{p;\alpha;Base}$ or Fisher's test or Chi-square test) as well as with hypotheses of a different nature. Let us give examples of further generalized quantifiers: for the parameters a...d, see Table 2.

Founded implication $\Rightarrow_{p;Base}$ with the parameters $0 < p \leq 1$ and $Base > 0$.

The condition $\frac{a}{a+b} \geq p \wedge a \geq Base$ is associated to quantifier $\Rightarrow_{p;Base}$.

The formula $\varphi \Rightarrow_{0,95;Base} \psi$ can be interpreted as "*95% of objects satisfying* φ *satisfy also* ψ" or "φ *implies* ψ *on the level of 95%*".

Double founded implication $\Leftrightarrow_{p;Base}$ with the parameters $0 < p \leq 1$ and $Base > 0$. The condition $\frac{a}{a+b+c} \geq p \wedge a \geq Base$ is associated to the quantifier $\Leftrightarrow_{p;Base}$, see also [Ullman, 2000].

The formula $\varphi \Leftrightarrow_{0,95;Base} \psi$ can be interpreted as "*95% of objects satisfying* φ *or* ψ *satisfy both* φ *and* ψ" or "$\varphi \wedge \psi$ *implies* $\varphi \vee \psi$ *on the level of 95%*".

Founded equivalence $\equiv_{p;Base}$ with the parameters $0 < p \leq 1$ and $Base > 0$.

The condition $\frac{a+d}{a+b+c+d} \geq p \wedge a \geq Base$ is associated to quantifier $\equiv_{p;Base}$.

The formula $\varphi \equiv_{0,95;Base} \psi$ an be interpreted as "*95% of objects have the same value for* φ *and* ψ".

The 'classical' associational rule [Aggraval, 1996] can also be understood as a generalized quantifier with the condition $\frac{a}{a+b} \geq C \wedge \frac{a}{a+b+c+d} \geq S$ associated to it where C is the confidence and S is the support (see Section 6.2.1).

A relation of two Boolean attributes is defined in [Zembowicz, 1996].

It can be understood as a generalized quantifier with $\approx^{E}_{\delta}$ with the condition $\frac{a}{a+b} < \delta \wedge c+d > 0 \wedge \frac{c}{c+d} < \delta$ associated to it.

Examples

A condition concerning the four-fold table is associated with each generalized quantifier. An example of the generalized quantifier is the quantifier $\Rightarrow^{!}_{p;\alpha;Base}$ of lower critical implication with the parameters $0 < p \leq 1$, $Base > 0$ and $0 < \alpha \leq 0.5$.

A condition $\sum_{i=a}^{r} \frac{r!}{i!(r-i)!} * p^{i} * (1-p)^{r-i} \leq \alpha \wedge a \geq Base$ is associated to the quantifier $\Rightarrow^{!}_{p;\alpha;Base}$. The relation $\varphi \Rightarrow^{!}_{p;\alpha;Base} \psi$ corresponds to a test (on the level α) of a null hypothesis H_0: $P(\varphi|\psi) \leq p$ against the alternative one H_1: $P(\varphi|\psi) > p$. If the condition associated to $\Rightarrow^{!}_{p;\alpha;Base}$ is true in the analyzed data matrix, then we say that the formula $\varphi \Rightarrow^{!}_{p;\alpha;Base} \psi$ is true in the analyzed data matrix and the alternative hypothesis is accepted.

Another example of the generalized quantifier example is Fisher's quantifier $\sim_{\alpha,Base}$ with parameters $0 < \alpha \leq 0.5$ and $Base > 0$.

The condition $ad > bc \wedge \sum_{i=a}^{\min(r,k)} \frac{r!s!k!l!}{n!i!(r-i)!(k-i)!(n-r-k-i)!} \leq \alpha \wedge a \geq Base$ is associated to the quantifier $\sim_{\alpha,Base}$. The relation $\varphi \sim_{\alpha,Base} \psi$ corresponds to a test (on the level α) of the null hypothesis of the independence of φ and ψ against the alternative one of the positive dependence.

Examples of hypotheses

– **age(35–40) ∧ sex(F) ∧ syst(>150) $\Rightarrow^{!}_{0.95;0.05;100}$ diag(D_Z),**
  or in other words:
  the combination of the properties age between 35 and 40 AND sex being female AND a systolic pressure of over 150 implies diagnosis Dz with a probability of 95% and a

significance level of 0.05 and there are at least 100 patients with diagnosis Dz satisfying the combination of the properties.

– **age(50–60) ∧ sex(M) ∧ diag(D_Z) ~$_{0.05,50}$ syst(>160) ∧ diast(>90).**
In other words:
the combination of the properties age between 50 and 60 AND sex being male AND being diagnosed Dz is associated with a systolic pressure higher than 160 in combination with a diastolic pressure higher than 90 and there are at least 50 patients satisfying both combinations. 'Is associated' means that the Fisher's test applied to combinations of properties in question is significant on level 0.05.

The LISP-miner system

LISP-Miner is an academic experimental software system developed for the Ssupport of research and teaching activities in knowledge discovery in data. M. Šimůnek and students of the Faculty of Informatics and Statistics of the University of Economics in Prague played an imported role in its development. The core of the LISP-Miner is the procedure 4ft-Miner [Rauch, 2000]. It has a rich variety of parameters for the definition of the set of relevant questions and various tools for the interpretation of the resulting tests of relevant assertions. Also a data preprocessing module called DataSource is integrated. Further, a new version of the machine learning procedure KEX [Berka, 1994] is being integrated. The last version of the system LISP-Miner is included on the CD-rom.

Tailor made interface

Various types of tasks can be solved by the application of 4ft-Miner, e.g. *What are possible causes of a given phenomenon?* and *Under which conditions are two attributes equivalent?*

Usually there are several options to solve a particular task. The task *What are possible causes of a given phenomenon?* can be solved using the generalized quantifier of the founded implication $\Rightarrow_{p;Base}$ or using the quantifier of the lower critical implication $\Rightarrow^{!}_{p;\propto;Base}$, various attributes and various types of basic Boolean attributes can also be used.

It can be very difficult to choose suitable parameters to define the set of relevant questions. Similarly it can be a difficult task to interpret the resulting set of relevant assertions. A high level of knowledge of the 4ft-Miner procedure may be required.

Tailor made software communicating with a user in his task-dependent terminology that will both prepare parameters for 4ft-Miner and interpret its results is being developed. The starting point for this development is the fact that both parameters defining the set of relevant questions and the resulting set of relevant assertions are stored in a well-defined database.

Thus, the parameters defining the set of relevant questions can be prepared by a program independent of 4ft-Miner and stored in this database. The module generating and testing relevant questions can read prepared parameters and work with them. Similarly the resulting relevant assertions can be read and further interpreted by a tailor-made program. The whole solution is sketched in Figure 3. The first version of the module for a tailor-made interface is also included on the disk.

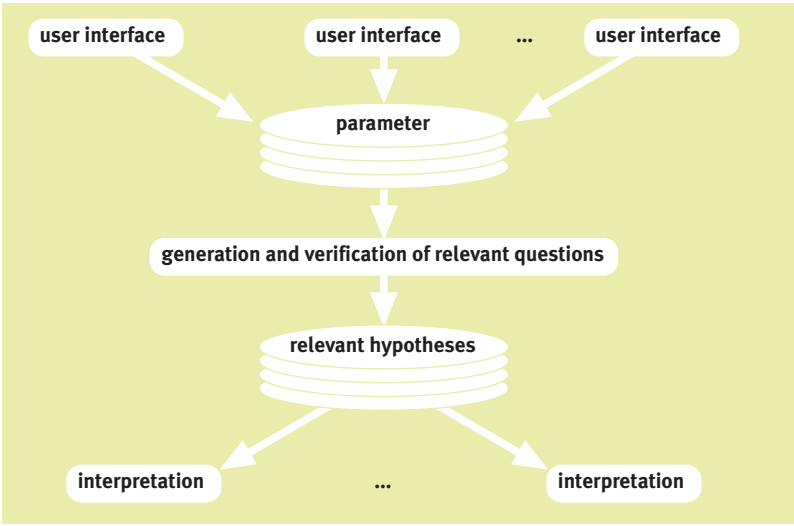


Figure 3
Tailor made interfaces to the 4ft-Miner procedure.

Further possibilities

Analytical reports and logical calculi

Experience shows that it is useful to arrange results of data mining into an analytic-synthetic report structured according to the analyzed problem. The core of such a report is a set of relevant hypothesis concerning the analyzed data (not necessarily in the form of Antecedent ≈ Succedent). Thus, we can understand the whole report as a finite set of formulas of an observational calculus.
Understanding the whole analytic-synthetic report as a finite set of formulas of the logical calculus offers some new possibilities. We can try to define a logically minimal skeleton of such a report. The logically minimal skeleton of the report will be set of formulas of a suitable logical calculus. It will be possible to deduce the whole report by formal deduction rules from the logically minimal skeleton. Appropriate correct deduction rules are necessary to solve this task.
Various types of analytic-synthetic report can be defined. A method for an automated production of the logically minimal skeletons of a particular type of report from the given data can be developed. Thus, special modules operating on large data warehouses can permanently produce particular analytical reports represented by logically minimal skeletons. Resulting reports can be offered via Internet, etc.

We can use the logically minimal skeleton of the report to represent its content in the same way that index terms are used to represent the content of a textual document in information retrieval. Unlike index terms in information retrieval, the logically minimal skeleton will describe the content of the report in a precise way. It will be possible to deal with the logically minimal skeletons instead of with the whole reports. It will be possible to do all kinds of manipulation with analytical reports by computer including manipulation with the content. The content will be represented by logical formulas and the computer will understand it by means of formal logic.
E.g. let us suppose we have a large set of analytical reports, each of them represented by the logically minimal skeleton. Thus, we can solve the task to find all reports dealing with a given problem in the same way as a given report. It is not possible to solve such a task using usual index terms.

An example of the deduction rule is the relation $\dfrac{\varphi_1 \Rightarrow^{!}_{0.9,\,0.05,\,100} \psi_1}{\varphi_2 \Rightarrow^{!}_{0.9,\,0.05,\,100} \psi_2}$

concerning two formulas $\varphi_1 \Rightarrow^{!}_{0.9,\,0.05,\,100} \psi_1$ and $\varphi_2 \Rightarrow^{!}_{0.9,\,0.05,\,100} \psi_2$.

This deduction rule is correct if the following condition is true: if the formula

$\varphi_1 \Rightarrow^{!}_{0.9,\,0.05,\,100} \psi_1$ is true in the given data matrix, then also the formula

$\varphi_2 \Rightarrow^{!}_{0.9,\,0.05,\,100} \psi_2$ is true in this data matrix.

Research shows that there is a simple way to decide if a given deduction rule is correct [Rauch, 1998a]. This depends on the class of the generalized quantifier. There are various classes of generalized quantifiers, e.g. the class of implication quantifiers or the class of double implication quantifiers.

Quantifier $\Rightarrow^{!}_{p;\alpha;Base}$ of the lower critical implication quantifier $\Rightarrow_{p;Base}$ of the founded implication are examples of implication quantifiers. Quantifier $\Leftrightarrow_{p;Base}$ of double founded implication $\Leftrightarrow_{p;Base}$ is the example of the double implication quantifier.

The example of a correct deduction rule is $\dfrac{\mathrm{sex(F)} \Rightarrow^{!}_{0.9,\,0.05,\,100} \mathrm{diag(D_A)}}{\mathrm{sex(F)} \Rightarrow^{!}_{0.9,\,0.05,\,100} \mathrm{diag(D_A, D_B)}}$,

deduction rule $\dfrac{\mathrm{age(35\text{ - }40)} \Rightarrow^{!}_{0.9,\,0.05,\,100} \mathrm{diag(D_A)}}{\mathrm{age(35\text{ - }40)} \Rightarrow^{!}_{0.9,\,0.05,\,100} \mathrm{diag(D_A)} \wedge \mathrm{sex(F)}}$ is not correct.

New GUHA procedures

It is typical for the implemented GUHA procedures that the input for one run of the procedure is a single data matrix. These procedures are not intended for the analysis of more complex data structures stored in databases. It is theoretically possible to transform each data structure stored in the database into a single data matrix. However, this approach is not practically efficient.

One reason is in the size of the data matrix resulting from SQL transformations. The second reason is that expressing interesting patterns by means of a complex database model is much simpler than expressing them by quantities corresponding to data matrix columns.

Some special observational calculi seem to be suitable for GUHA procedures working on complex data structures stored in a relational database. More on these calculi can be found in [Date, 1976; Rauch, 1986].

A new version of the procedure 4ft-Miner is being implemented. This version is based on the suggestions given in [Rauch, 1986]. Simply speaking, the new version of 4ft-Miner analyses several related data matrices forming a tree structure.

Let us point out that further GUHA procedures are suggested in [Hájek, 1978]. Examples are procedures based on Spearmen's rank correlation coefficient and on Kendall's rank correlation coefficient. Further GUHA-like analytical procedures can work with an n x m contingency table instead of a four-fold contingency table.

This work is supported by grant EU No IST-1999-11495, Data mining and decision support for business competitiveness, Solomon European Virtual Enterprise.

References

- Aggraval, R., et al. (1996). Fast Discovery of Association Rules. In: U.M. Fayyad, et al. Advances in Knowledge Discovery and Data Mining. AAAI Press, The MIT Press. pp307-328
- Berka, P., J. Iválnek. (1994). Automated Knowledge Acquisition for PROSPECTOR-like Expert Systems. In: Bergadano, deRaedt (eds.). Proceedings ECML'94. Springer Verlag
- Date, C.J. (1976). An Introduction to Database Systems. Addison-Wesley Publishing Company
- Hájek, P., T. Havránek. (1978). Mechanising Hypothesis Formation - Mathematical Foundations for a General Theory. Springer Verlag
- Hájek, P., T. Havránek, M. Chytil. (1983). Metoda GUHA. Praha, Academia. (in Czech)
- Hájek, P., A. Sochorová, J. Zvárová. (1995). GUHA for Personal Computers. Computational Statistics & Data Analysis **19**:149-153

- Hájek, P., M. Holeňa. (1998). Formal Logics of Discovery and Hypothesis Formation by Machine. In: S. Arikawa, H. Motoda. (eds.). Discovery Science. Springer Verlag. pp291-302
- Iványek, J. (1999). On the Correspondence between Classes of Implicational and Equivalence Quantifiers. In: J. Zytkow, J. Rauch. (eds.). Principles of Data Mining and Knowledge Discovery. Berlin, Springer Verlag. pp.116-124
- Rauch, J. (1978). Some Remarks on Computer Realizations of GUHA Procedures. International Journal of Man-Machine Studies **10**:23-28
- Rauch, J. (1986). Logical Foundations of Mechanising Hypotheses from Database. PhD. Thesis. Mathematical Institute of Czechoslovak Academy of Sciences, Prague. (in Czech)
- Rauch, J. (1997a). Logical Calculi for Knowledge Discovery in Databases. In: J. Komorowski, J. Zytkow. (eds.). Principles of Data Mining and Knowledge Discovery. Springer Verlag. pp47-57
- Rauch, J. (1998a). Classes of Four-Fold Table Quantifiers. In: M. Quafafou, J. Zytkow. (eds.). Principles of Data Mining and Knowledge Discovery. Springer Verlag. pp203-211
- Rauch, J. (1998b). Four-fold Table Calculi and Missing Information. In: P. Wang (ed). JCI'S98 Association for Intelligent Machinery. Vol. II. Duke University, Durham
- Rauch, J., M. Šimůnek. (2000). Mining for 4ft Association Rules. In: S. Arikawa, S. Morishita. (eds.). Discovery Science 2000. Springer Verlag. pp268-272
- Ullman, J. (2000). A Survey of Association-Rule Mining. In: S. Arikawa, S. Morishita. (eds.). Discovery Science 2000. Springer Verlag. pp1-14
- Zembowicz, R., J. Zytkow. (1996). From Contingency Tables to Various Forms of Knowledge in Databases. In: U.M. Fayyad, et al. Advances in Knowledge Discovery and Data Mining. AAAI Press, The MIT Press. pp329-349

2.2.5 Data access for atmospheric research

Wim Som de Cerff[1], *John van de Vegte*[2], *Richard M. van Hees*[3]

Introduction

As data acquisition techniques advance, enormous quantities of data are gathered and stored. Often this storage is located at a great distance from the actual location of the scientists. Furthermore storage space itself is distributed geographically. These circumstances require special solutions to ensure scientific access to the data. This article will focus on this matter from the perspective of atmospheric research.

The measurements of earth observation instruments are important input for climate and weather models and are used for research on atmospheric composition and climate. These instruments, mounted on earth orbiting satellites, produce Tera to Peta byte archives, distributed over several archiving sites.
This introduces two problems for the individual scientist: How to find the information needed in these vast archives and, if the data of interest is found, how to process these Tera bytes of data into usable parameters?

Earth observation data access

Several satellites containing atmospheric research instruments are orbiting the earth or will be launched in the near future. This paper will only mention three of these instruments: GOME, SCIAMACHY and OMI, this is because of the involvement of SRON and KNMI in these instruments.

Data properties

The GOME (Global Ozone Monitoring Experiment) is mounted on the ERS-2, which was launched in 1995 by ESA (European Space Agency). The instrument is still in operation and will be followed up by SCIAMACHY (SCanning Imaging Absorption SpectroMeter for Atmospheric CartograpHY), mounted on the ESA-ENVISAT (ESA-ENVIronmental SATellite), which will be launched in 2002. The SCIAMACHY primary mission objective is to perform global measurements of trace gases (like ozone) in the troposphere and stratosphere. OMI (Ozone Monitoring Instrument) will be mounted on the NASA-AURA spacecraft and will be launched in 2004. All instruments measure sun absorption and or reflection from which different trace gases can be derived. Figure 1 shows the measurement spectrum of SCIAMACHY. The horizontal axis shows the wavelength measured. The vertical axis shows the attenuation. The trace gases derived from the measurement are named at the top of their peaks in the spectrum.

1 W.J. Som de Cerff, sdecerff@knm.nl, Royal Netherlands Meteorological Institute (KNMI), Observations and Modelling Department, Satellite Data Division (WM/SD), De Bilt, The Netherlands

2 J. van de Vegte, vegtevd@knmi.nl, Royal Netherlands Meteorological Institute (KNMI), Observations and Modelling Department, R&D Observations Division, De Bilt, The Netherlands, http://neonet.knmi.nl

3 R.M. van Hees, Space Research Organisation Netherlands (SRON), Utrecht, The Netherlands

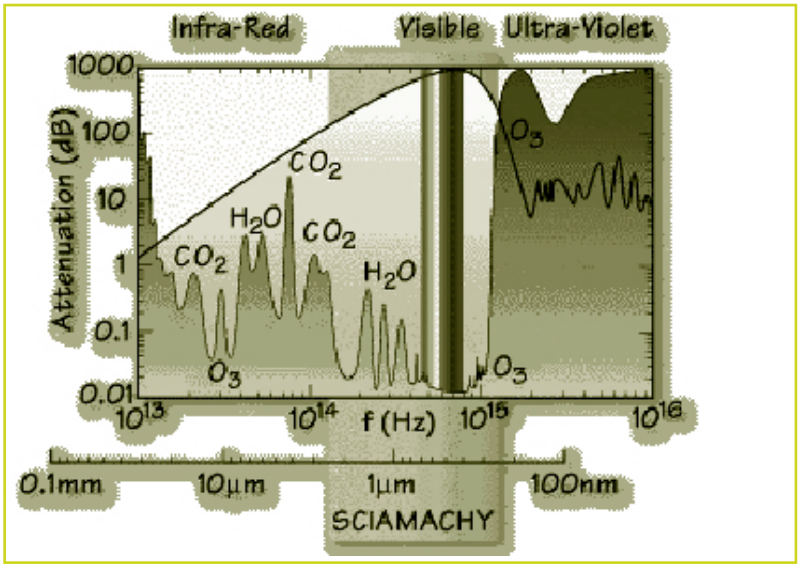


Figure 1
Measurement spectrum of SCIAMACHY.

The data rate of the earth observation instruments is growing with each new generation. The GOME instrument has a data rate of 80 Mbyte per day, SCIAMACHY will have a data rate of 3 Gbyte per day and the OMI instrument will have a data rate of 41 Gbyte per day. All instruments will measure for (at least) five years. From these rates data products will be processed. This will lead to multi Petabyte data archives, distributed over several archives.
In earth observation data is classified into levels, according to how much processing has been performed on the data. Measurements from the instrument, the 'raw' measurement as observed by the space-borne instrument, are called level 0 data. This level 0 data is processed into level 1 data, which is time referenced and geo-located. Level 1 data is processed into level 2 data, which contains the derived geophysical parameters (e.g. O_3, CO_2) at the same spatial resolution as level 1 data. Level 2 data can be processed further to level 3 or level 4 data, where the data is re-gridded or combined with other data sources, such as model output or data from other instruments. Different products (at level 2, 3 or 4) can be derived from the level 1 product.

The processing of the satellite instrument measurements into data products which can be used for research is a computing intensive operation. To retrieve , for instance, an ozone profile level 2 product from one orbit [4] of level 1 data takes respectively 2 minutes for GOME, 2,000 minutes (1,4 days) for SCIAMACHY and 320,000 minutes (222,2 days) for OMI, based on processing on one MIPS R10000 250 MHz processor.
Data archives are also distributed over several locations. GOME and SCIAMACHY level 1 and 2 data are located at DLR (Germany). OMI level 1 and 2 data will be archived at the Distributed Active Archiving Centre (DAAC, Goddard

4 One orbit is approximately 100 minutes of data.

Space Flight Centre, Washington D.C. USA). Higher level products are scattered over all kinds of locations (including KNMI and SRON).

Challenges

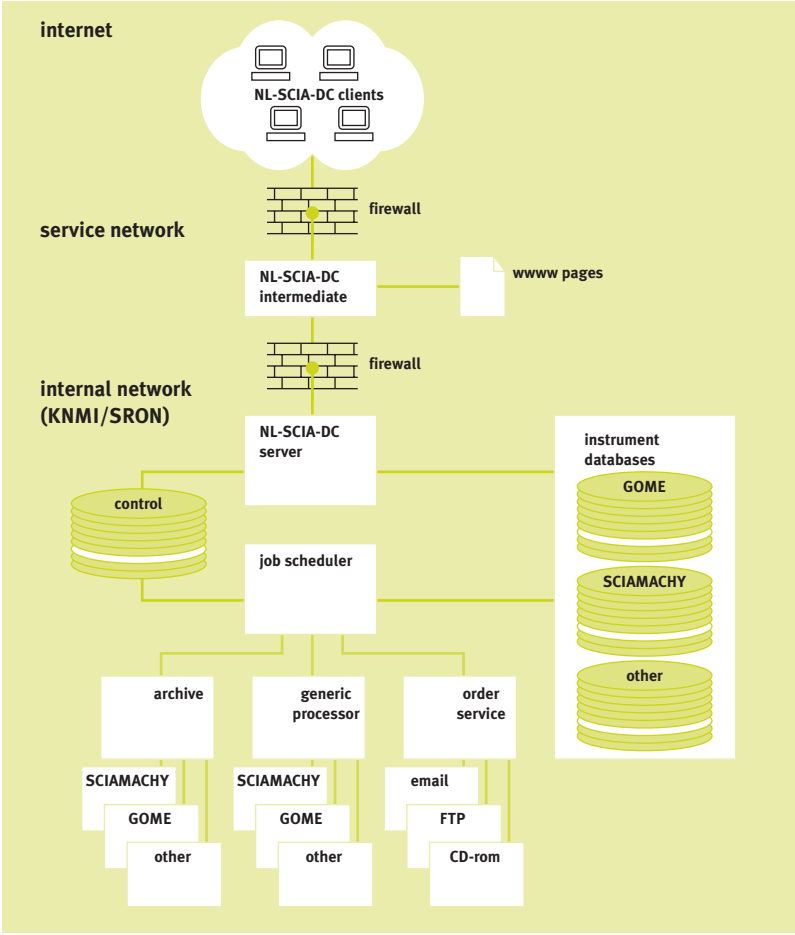
The amount of storage and computing required exceeds the available resources of individual scientists. Therefore the data archiving and processing of derived products must be organized in such a way that the scientist can easily find the data and retrieve information of interest. The use of meta-data, visualization, distributed processing (sharing processing capacity) and interactive queries support the process of finding the necessary information. Another challenge provides remote access to these distributed databases.

Interfacing and mining: the NL-SCIA-DC project

The Netherlands SCIAMACHY Data Center (NL-SCIA-DC) aims at providing access to measurements from GOME and SCIAMACHY (both earth observation instruments mounted on satellites). The NL-SCIA-DC provides data selection, processing and downloading mechanisms to scientists.
The need for the NL-SCIA-DC came from the atmospheric researchers in the Netherlands who needed faster access to and more flexibility in accessing and processing GOME data.
The current NL-SCIA-DC version is an Internet-enabled application, which is used by the science community to select, process and download GOME data. It also contains facilities to test new data processors. The core of the system is the control database (Figure 2). Here the users, instrument data, data processors and their parameters, data processor locations and the locations of other databases are administrated. The control database is used for building the graphical user interface (GUI) of the NL-SCIA-DC and for controlling the process and download system. The fields are displayed in the GUI, so that the user can see the status of the jobs submitted.

The GUI is built dynamically [Vegte, 2000] from settings in the control database. This allows new processors and new databases to be integrated easily. The GUI will automatically adapt to the new settings. The development of a dynamic interface takes more time then hard coding all the options directly in the GUI. However this investment will pay itself back in a more flexible system allowing easier adaptation to new demands, less maintenance and higher availability of the data center.
The data center offers three ways of selection: catalogue, browse and query. The catalogue style is a file-manager way of selecting whole files of data. The browse selection allows a graphical method of selection: satellite tracks can be displayed on a world map. This allows the user to select an interesting scene and download the corresponding data.

Figure 2
System layout.



The third way is a query on the available (meta) data parameters of the data (e.g. ozone value, zenith angle, pixel type). The NL-SCIA-DC is accessible through the NL-SCIA-DC web site[5].

Scientists have shown their interest in using the NL-SCIA-DC. Currently the site has registered users from the Dutch GOME science community and other institutes (Eumetsat, ESA-ESTEC, BIRA, University of Leicester). New research projects have started which will use the NL-SCIA-DC as basic tool for their research.

Remote access: the DataGRID project

Grids are the computer networks of the future. Grids will, in analogy with the electricity networks, eventually cover the world and give the user uncomplicated access to both (super) computing facilities and data banks all over the globe. The DataGRID project is a European project, lead by CERN, in which 23 partners from different disciplines (High Energy Physics, Earth Observation and BioScience) are cooperating. Their main goal is to enable next generation scien-

5 http://neonet.knmi.nl/neoaf/

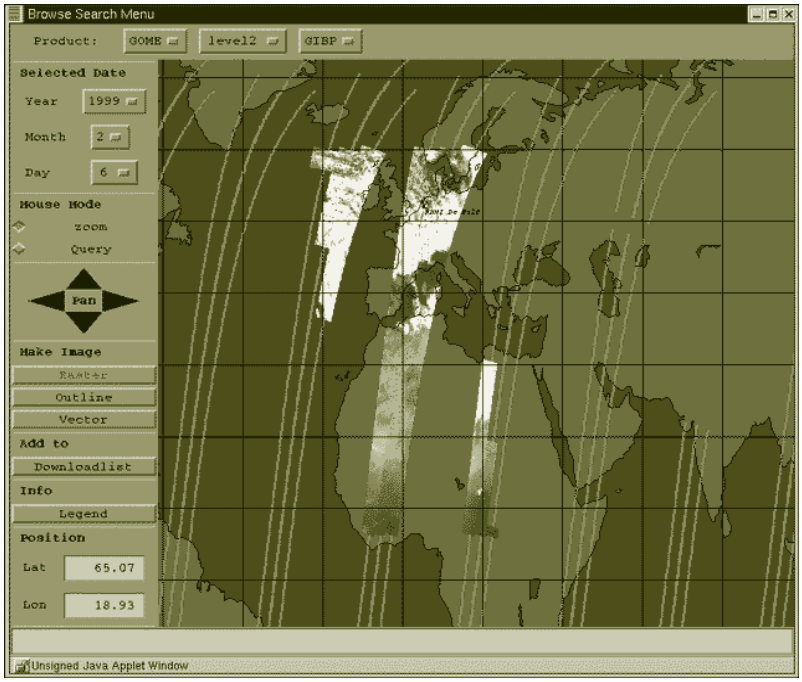


Figure 3
Browse window with visualization of satellite tracks.

tific exploration, which requires intensive computation and analysis of shared large-scale databases, from hundreds of TeraBytes to PetaBytes, across widely distributed scientific communities.

We see these requirements emerging in many scientific disciplines, including physics, biology, and earth sciences. Such sharing is complicated by the distributed nature of the resources to be used, the distributed nature of the communities, the size of the databases and the limited network bandwidth available. To address these problems emerging computational Grid technologies are used as a basis to:

- establish a research network that will enable the development of the technology components essential for the implementation of a new worldwide Data Grid on a scale not previously attempted;
- demonstrate the effectiveness of this new technology through the large-scale deployment of end-to-end application experiments involving real users;
- demonstrate the ability to build, connect and effectively manage large general-purpose, data intensive computer clusters constructed from low-cost commodity components.

These goals are ambitious. However, by combining recent research results from and collaborating with other related Grid activities throughout the world, this project can focus on developments in the areas most affected by data organization and management. (Quoted from the DataGRID project proposal [DataGRID, 2000]

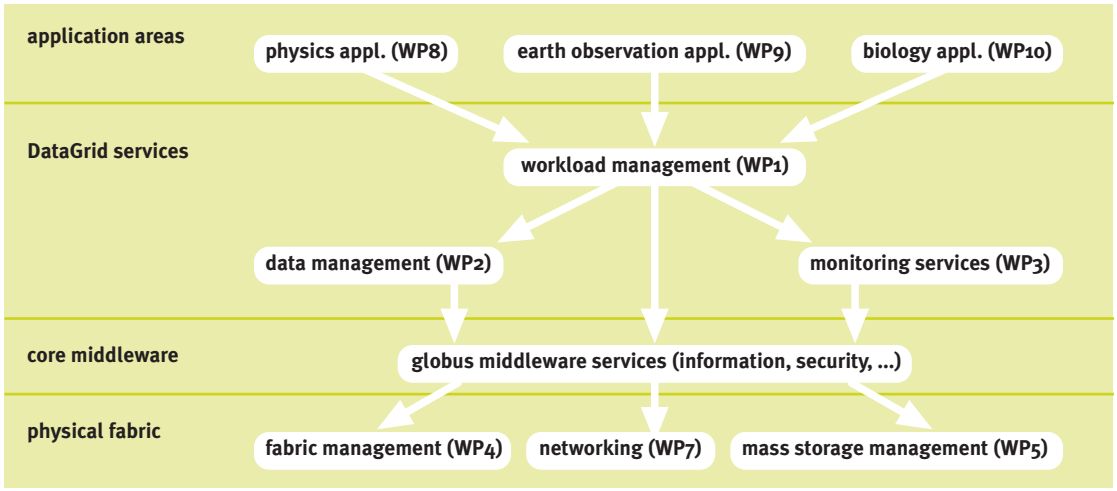


Figure 4
DataGRID project structure.

The project is divided in several Work packages (WP) and will build on the Globus Middleware[6]. The Work packages are spread over four areas, as depicted in Figure 2. The application areas will build test applications which will use the DataGRID services.
The DataGRID Project is a proposal approved by the European Commission for shared cost research and technological development funding. The project has six main partners (CERN, CNRS, ESA, INFN, NIKHEF and PPARC) and fifteen associated partners, among which are SARA and KNMI.
The KNMI is involved in the Earth Observation Work package (WP9). The aim is to help define the requirements and provide cases of DataGRID usage to the Middleware Work packages, but the main activity is the creation of an application, running on the GRID. This application will be the retrieval of ozone profiles from GOME and, later on, SCIAMACHY data. Within the DataGRID project there is cooperation between the Dutch partners in the project DutchGrid.

DutchGrid[7] implements a small-scale test bed Grid infrastructure based on the Globus toolkit. The initial deployment concerned both institutes on the WCW campus (NIKHEF, the UvA Computer Science Institute, AMOLF, CWI, and SARA), and our EU DataGRID partner institute KNMI. The original foundation for the DutchGrid test bed was deployed as part of the 'ICES-KIS' Virtual Laboratory project (now called VLAM-G).

The basic component of the DutchGrid infrastructure is a shared security domain, centered around the DutchGrid/NIKHEF Certification Authority. Any Dutch research institute is welcome to use the DutchGrid CA for its test bed research and participation even at this early level is quite welcome. Note that this does not imply access to your resources by unknown third parties: authorization and authentication are clearly separated concepts in DutchGrid

6 http://www.globus.org

7 http://vlabwww.nikhef.nl

[DutchGrid, 2001]. The DutchGrid will be used for early on testing of the Ozone profile application. The sharing of knowledge between the Dutch partners has proven to be fruitful.

Perspective

The Goal of the KNMI and SRON is to combine the results of both projects into one portal functionality: by using the NL-SCIA-DC interface, a researcher can select the data of interest, and submit a processing request, regardless how much data is selected or where the data is located. The researcher can then use the results for visualization (2D or 3D) or as input for (climate) models, which in their turn can be executed in the Grid. In this way, the researchers will have access to all kinds of data, can process data using their new or experimental processors and can publish their results to the community. In Figure 4 this view is depicted. But before this view can become reality, some problems have to be tackled.

Bandwidth

First of all, the network capacity of all institutes needs to be sufficient in order to be able to use the various processing sites and data archives. In the Netherlands GigaPort (SurfNET 5) will provide the Dutch academic world with a new fast network, where all DutchGrid partners will (at least) have Gigabit connection to each other. In Europe the Dante project aims at connecting all European research networks. Across the Atlantic other initiatives are in place to upgrade connectivity.

Global standards

The development of Grid software is a global effort as well. Without some kind of standardization there cannot be a global usable Grid. For this reason the Global Grid Forum[8] (GGF) has been founded. GGF focuses on the promotion and development of Grid technologies and applications via the development and documentation of 'best practices', implementation guidelines, and standards with an emphasis on rough consensus and running code. Global GF participants come from over 200 participating organizations in over 30 countries.

Processing and visualization

The NL-SCIA-DC is a good starting point for the development of an atmospheric portal providing functions demanded by the atmospheric research community. At this moment, it already provides some of the basic functionality. But the demands of the atmospheric research community go beyond what is offered now.

Processing is based on processing on file level, i.e. query or browse results cannot be processed. Researchers use the query and browse options to select files

8 http://www.gridforum.org

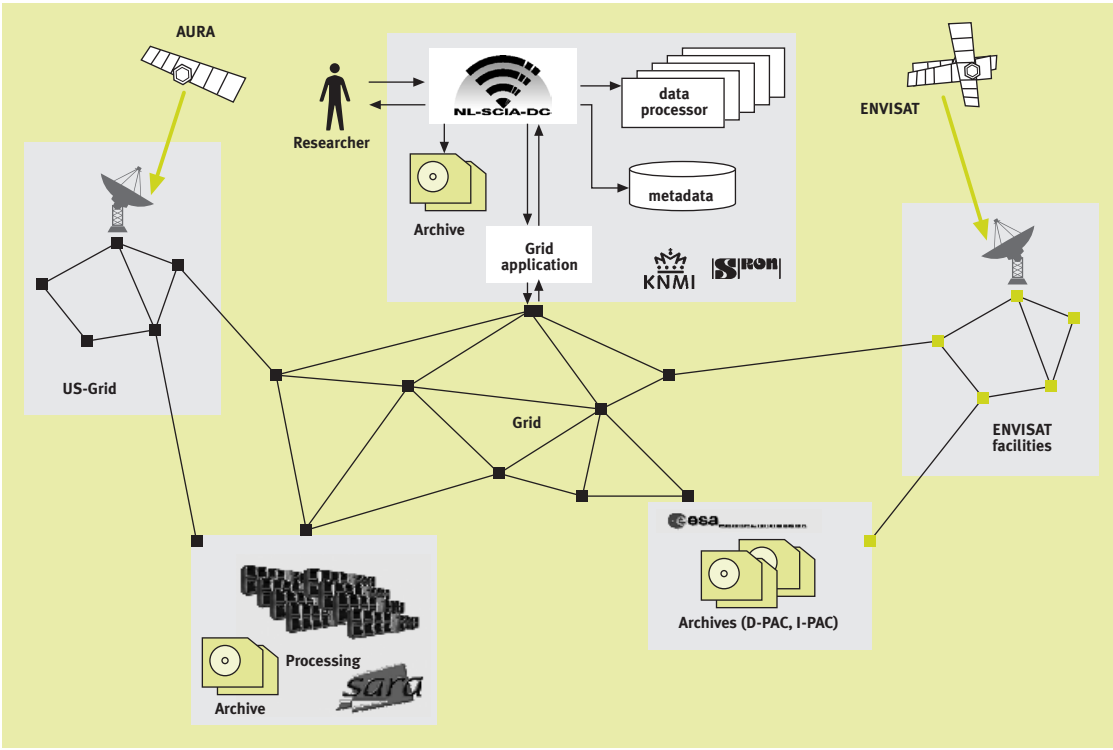


Figure 5
View of a Grid application: The Grid can be used to collect the data directly at the different archives of ESA and NASA. The KNMI can make their own products (using the ESA/NASA data) which can be made available to users of the Grid and to other users (using the NL-SCIA-DC). The Grid can be used to process these products, i.e. process job is started at the NL-SCIA-DC. This job needs data located at the ESA archives. The job collects the required data and transports itself to SARA where it starts processing. The product is transported back to the NL-SCIA-DC where it is used.

and perform their processing or analyzing efforts on those files. For example, the researcher searches for cloudless scenes using the Query menu or the Browse menu. The search result is an ASCII table, but currently no further processing can be done using this file. Also the visualization is limited to 2D plotting, while the researchers would like 3D visualization. When data from the new instruments (OMI, SCIAMACHY) becomes available, researchers also want to be able to do cross comparisons of the measurements. They also want to compare their measurements with measurements from other sources (e.g. model output, ground based measurements or measurements from other satellite instruments).

Other sciences

SRON and KNMI and the Earth Observation community are not the only institutions which will see their archive and computational needs grow dramatically in the coming years. There are other research fields which will even have higher needs (e.g. High Energy Physics, Astronomy, Bio-sciences). Grid technology has just started to provide helpful solutions to the data and computational problems. A lot of (large) Grid related projects have started or will start in the near future (e.g. UK E-science project, AstroGRID, IGRID). Especially the collaborative science projects, aimed at providing Grid solutions to research groups so that they can combine their research efforts, will provide helpful solutions.

Conclusion

How can the information needed be found in the huge archives and, when data of interest is found, how can these vast data amounts be processed into usable parameters? These questions are only partly addressed by the developments of DataGRID and NL-SCIA-DC. New projects, building on the results of these two projects, will provide an adequate answer in the future. When researchers want to obtain the maximum result from the very costly space borne instruments, these future projects are a prerequisite.
The combination of the two projects can provide an approach for a solution for atmospheric research on huge data sets. The NL-SCIA-DC provides the interface to the (meta) data and the data mining functionality, while at the back-end DataGRID technology can be used to deal with the data location and computation problems.

The solution sketched above may provide a solution for a limited number of researchers, but it is no general solution for all data and computing problems. Collaborative science (E-science) implementations will provide more general solutions in the future.

References

- Kroll, M. et al. (1999). User Requirements Document V1.1. NL-SCIA-DC-URD-1.1. http://neonet.knmi.nl/neoaf/reports.html
- Som de Cerff, W.J., J. van de Vegte, R.M. van Hees. The Netherlands SCIAMACHY Data Center. 2nd International Symposium on Operationalization of Remote Sensing. Enschede, The Netherlands
- Vegte, J. van de, et al. (2000). Dynamic GUI for Accessing the Netherlands. SCIAMACHY Data Center. DASIA 2000. ESA SP457. Montreal, Canada
- Barlag, S., J.A. Beysens, R.M. van Hees, R.B.A. Koelemeijer, A.J.M. Piters, H. Schrijver, P. Stammes, W.J. Som de Cerff, P. Valks, J van de Vegte. (2000). Design of the Netherlands SCIAMACHY Data Center. ERS - ENVISAT Symposium 2000. Gothenborg, Sweden
- Foster, I., C. Kesselman, et al. (1999). The GRID, a Blueprint for a New Computing Infrastructure. Morgan Kaufmann Publishers
- Research and Technical Development for an International Data Grid. DataGRID description of work. IST-2000-25182
- NEONET program. http://www.neonet.nl or http://apex.neonet.nl
- The NL-SCIA-DC. http://neonet.knmi.nl
- Dutch DataGRID. http://www.datagrid.nl

2

2.3 Application in Science Areas

2.3.1 Knowledge discovery in medical databases

Shusaku Tsumoto[1]

Introduction

Medical progress is commonly supported by data analysis, which not only discovers diseases, but also improves the skill of medical experts and establishes the treatment technique of diseases, although the techniques used have been mostly manual or semi-manual. One of the illustrative examples is epidemiological research of chronic diseases [Rothman, 1998]. After statistical approaches had been established, the application of statistics found from large databases that hypertension and cholesterol are closely related with cerebral stroke and ischemic heart diseases, which resulted in the development of a large number of chemicals for control of hypertension. It is already widely accepted that these chemicals contribute to the decrease in the prevalence of cerebral stroke.

1 Prof S. Tsumoto, tsumoto@computer.org, Department of Medical Informatics, School of Medicine, Shimane Medical University, 89-1 Enya-cho Izumo City, Shimane 693-8501, Japan

It is true that the development of medicine is not only supported by data analysis. Both medical practice and research have been changed by the rapid growth of life sciences, including biochemistry and immunology [Levinson, 1996]. The mechanism of a disease can be explained as a biochemical process or cell disorder and the diagnostic accuracy of medical experts is increasing due to the development of laboratory examinations. However, it is also true that data analysis is indispensable to generating a hypothesis. For instance, discovery of the HIV infection and Hepatitis type C were inspired by analysis of clinical causes unexpected by experts on immunology and hepatology [Fauci, 1997].

Although life science has advanced rapidly, the mechanisms of many diseases are still unknown: neurological diseases especially were very difficult to analyze, because their prevalence is very low [Adams, 1993]. Even the mechanisms of diseases with high prevalence, such as cancer are only partially known to medical experts. In this sense, medical research always needs a good hypothesis, which is one of the most important motivations to data mining and knowledge discovery for medical professionals.

Besides that medical researchers are interested in other aspects of data mining. Since the early 1980's, the rapid expansion of hospital information systems (HIS) has led to the storage of large numbers of laboratory examinations as databases [Bemmel, 1997]. For example, in a university hospital, where more than 1,000 patients visit from monday to friday, a database system stores more than 1 GB numerical data of laboratory examinations for each year. Furthermore, the storage of medical images and other types of data are discussed in medical informatics as research topics; electronic patient records and all the medical data will be stored in hospital information systems within the 21st century.

Thus, it is highly probable that data mining methods will find interesting patterns from databases, reusing stored data. These methods are important for medical research and practice, because human beings cannot deal with such a huge amount of data.

Problems of medical researchers

Medical researchers or practitioners would like to extract hypotheses which will lead to good medical research or medical practice. Research and practice do not always overlap, but good medical research will result in medical practice. Also, medical practice overlaps with healthcare which will have the same social objective. Thus, we will limit our discussions to issues on medical research in this article, although these are applicable to medical practice.

Data owners are interested in the following five problems:

1 Relations between examinations.
2 Searching for important factors for diagnosis.
3 Searching for important factors for prognosis.
4 Discovery of new diseases.
5 Short-term effects and long-term effects of therapeutic procedures.

The issues 2, 3 and 4 are closely related with the first one, so in the subsections below we mainly discuss the issues 1 and 5.

Relations between examinations

The final differential diagnoses are often made by the combination of specific examinations, such as medical image (CT, Computer Tomography or MRI, Magnetic Resonance Image) and immunological examinations. For example, introduction of CT has changed the quality of diagnosis of cerebral stroke [Adams, 1993]. Thus, people tend to think that the importance of traditional techniques has decreased in medical practice. However, conventional techniques are still important for medical practice for the following reasons.

Costs of medical image- and immunological tests

Medical image and laboratory examinations are much more expensive than physical examinations and they cannot be applied to each patient so often. Thus, if the specificity and sensitivity of a combination of physical examinations are equal to those of expensive tests, then the combination will be more applicable and useful.

Estimation of clinical courses and prognosis

For the differential diagnosis of consciousness loss, medical imaging such as CT and MRI, as well as laboratory examinations such as concentration of NH_3, are very important. However, in general they are not so useful for the determination of clinical causes and prognosis. Even now, the combination of physical examinations and medical expertise is useful to detect the status of each patient with a loss of consciousness [Plum, 1992]. The construction of prognostic models is one of the interesting topics in artificial intelligence in medicine [Lucus, 1999], although good models are very difficult to construct.

Discovery of patterns for syndromes

Although diseases with high prevalence are well examined and their precise classification is established, many diseases with low prevalence are not precisely classified and still described as syndromes, which are defined as a set of manifestations of which the pathophysiological processes are unknown [Harris, 1984; Adams, 1993]. Such syndromes may be classified into subclasses of the

same etiology[2]: the precise classification of low-prevalence diseases is required for future research, which is a strong motivation for data mining. It is also expected that such classification may lead to the discovery of new diseases.

Short-term effects and long-term effects

For the evaluation of treatments, statistical techniques are applied to indicate whether or not some chemicals are effective in treating a disease [Altman, 1991]. The time interval for analysis is at most five years, and most of the results are on short-term effects of therapeutic procedures. However, as the treatment of infectious diseases and chronic diseases has been established, most of the people keep healthy until they are 70 to 80 years old (in developed countries), which makes us realize the importance of the studies on long-term effects. For example, in the case of rheumatic and collagen diseases, oral steroid therapy has become the main therapy to suppress the activities of auto-immune disorders. However, several side-effects will be observed when this medication is taken for a long time and it is a very important research issue to estimate when such side-effects will start.

Although these studies on long-term effects are difficult when medical experts store and maintain their data by themselves, hospital information systems will enable them to store all the data much easier. Thus, it is expected that data mining techniques will contribute much to the analysis of data stored over long periods of time.

Available data

Two types of databases are available in the medical domain. One is a dataset acquired by medical experts, collected for a special research topic. For example, if a neurologist is interested in meningitis, he may start to collect all the symptoms and laboratory tests with some hypothesis in mind. These data have the following characteristics: 1. The number of records is small. 2. The number of attributes for each record is large, compared with the number of records. 3. The number of attributes with missing values is very small. These tendencies are general in scientific data, as discussed in [Westfall, 1993]. We refer to this type of database as prospective database or *p-database*. The analysis of those data is called *prospective analysis* in epidemiology [Kleinbaum, 1982; Rothman, 1998], because data collection is triggered by the generated hypothesis. Statistical analysis was usually applied to these datasets [Altman, 1991].

The other type is a huge dataset retrieved from hospital information systems. These data are stored in a database automatically without any specific research purpose. Usually, these databases only include laboratory tests, although researchers in medical informatics are discussing how to store medical images and physical examinations as electronic patient records [Bemmel, 1997].

2 Medical causes.

Especially, the standards of data types and their storage will be determined by ISO/TC215. After this establishment, these different types of data will be stored in HIS, available for data analysis. These data in HIS have the following characteristics: 1. The number of records is huge. 2. each record has a large number of attributes (more than several hundred). 3. Many missing values are present. 4. Many temporal subrecords are stored for each record (patient). We refer to this type of database as retrospective database (*r-database)*. The analysis of those data is called *retrospective analysis* in epidemiology, because data will be analyzed after data collection. Those data will lose many good features that prospective data holds and even statistical techniques will not perform well. This type of data is very similar to business databases, and this is where data mining techniques will be very useful.
The classification of data is very important for the analysis of medical databases. When readers are planning to analyze medical data, they should clarify which type of data they will analyze. In the subsequent sections, each discussion will be made for both database types.

Security problems

Concerning p-databases, medical experts are responsible for their security. Since medical experts are very sensitive to the security of information about their patients, p-databases will not include any personal information except for gender and age.

As for r-databases, security problems are still under discussion. Although almost all the HIS support several security measures, the level of security is variable and dependent on the interest of a system manager. Even researchers in medical informatics feel that the security of HIS is not sufficient. In ISO/TC215, this topic is very important and standardization of security problems is still being discussed in WG4, the results of which will be available in 2001. The discussion focuses on how to make the standards, that will be implemented in HIS within several years.

Data mining problems

Concerning p-databases, data will be prepared with a hypothesis, very carefully generated by medical experts. Thus, the quality of data is very high, and any data analysis technique will be applicable and useful. The problem with p-databases is that the number of measurements is very large, compared with the number of records. Thus, data reduction or rule induction will be useful to detect the important attributes for analysis. (See Sections 2.3.11, 6.2.4 and 6.2.7).
On the other hand, as for r-databases, there are many difficult issues for data analysis. In the following subsections, three important problems will be discussed.

Missing values

We can distinguish the following three types of missing values.

Occasional effects

This type of missing values is very often observed in business databases. Since data collectors may not have clear hypotheses for data collection, several attributes will not be recorded even though these attributes might be important. Thus, it is difficult for data miners to decide whether such attributes with missing values are important or not. Data analysis can be applied to such attributes, if the importance of these attributes cannot be checked during the preprocessing procedure.

Missing values concerned with medical decisions

Medical experts usually select physical and laboratory examinations, when they make a differential diagnosis in order to reach the final diagnosis as fast as possible. If medical experts decide that some examinations are not useful, they will not select such tests and attributes and the values of these tests will be stored as blank. In this case, missing values mean that these attributes are considered not important for a medical decision. Thus, data miners should ask data owners if a database has such attributes.

Progress in laboratory examinations

Many laboratory tests are developing, enabling us to gain accuracy of diagnosis and treatment. However, from the viewpoint of data analysis, these tests make data mining more difficult, because records taken before the introduction of a test will not include any information about this test. Thus, data miners should check whether a database includes such laboratory examinations.

Data miners should classify attributes with missing values and proceed into preprocessing of data. In the first type, conventional techniques for missing values can be applied (and removal of these attributes should be postponed). In the second type, context behind the missing values should be extracted and attached to the induced results. Finally, in the third case, data miners should change the strategy. Databases may be split into two sub-databases: one before the introduction of a new examination and the other one after the introduction, and data mining techniques should be applied to each sub-database.

Data storage

Medical examinations and observations are variable and sometimes difficult to be stored in fixed number of attributes. The following two types of data storage have to be preprocessed.

List of multivalued attributes

Some of the attributes should be represented as a list. For example, traffic accidents may injure several parts of bodies. Some patients have the damage only on hands and other ones suffer from multiple injuries, which makes it difficult to fix the number of attributes. If we enumerate all the possibilities of injuries and fix the number of columns corresponding to the worst case, most of the patients may require only a small number of columns. Usually, medical experts are not good at estimation of possible inputs, and they tend to make a list for data storage for the worst cases, although the probability for such cases is very low. For example, if medical experts empirically know that the number of injuries is at most 20, they will set up 20 columns for input. However, if the average number of injuries is 4 or 5, all the remaining attributes will be stored as blank. Although these attributes look like missing values, they should not be dealt with as missing values and have to be preprocessed: such large columns should be transformed into binary ones. For the above example, each location of injury will be appended as a column, and if that location is not described in a list, then the value of that column should be set to 0.

Temporal data

The characteristics of medical temporal databases are: 1. Each record is inhomogeneous with respect to time-series, including short-term effects and long-term effects. 2. Each record has more than 1,000 attributes when a patient is followed for more than one year. 3. When a patient is admitted for a long time, a large amount of data is stored in a very short term.

Since incorporating temporal aspects into databases is still an ongoing research issue in database science [Abiteboul, 1995], temporal data are generally stored as a table in hospital information systems (HIS). Table 1 shows a typical example of medical data, which is retrieved from a HIS. The first column denotes the ID number of each patient, and the second one denotes the date when the examinations took place that provided the datasets in this row. Each row with the same ID number describes the results of laboratory examinations, which were taken on the date in the second column. For example, the second row shows the data of the patient ID 1 on 04/19/1986.

We can list several other characteristics of medical temporal databases:

1 *Too many attributes*. Even though the dataset of a patient focuses on the transition of each examination (attribute), it would be difficult to see its trend when the patient is followed for a long time. If one wants to see a long-term interaction between attributes, it would be almost impossible. In order to solve this problem, most of HIS systems provide several graphical interfaces to capture temporal trends [Bemmel, 1997]. However, the interactions

ID	Date	GOT	GPT	LDH	γ-GTP	TP	Edema	···
1	19860419	24	12	152	63	7.5	–	···
1	19860430	25	12	162	76	7.9	+	···
1	19860502	22	8	144	68	7.0	+	···
1	19860506							···
1	19860508	22	13	156	66	7.6	–	···
1	19880826	23	17	142	89	7.7	–	···
1	19890109	32					–	···
1	19910304	20	15	369	139	6.9	+	···
2	19810511	20	15	369	139	6.9	–	···
2	19810713	22	14	177	49	7.9	–	···
2	19810907	22	12	173	50	7.6	–	···
2	19811102	34	46	159	104	7.3	–	···
2	19811130	16	6	161	44	7.1	–	···
2	19880826	23	17	142	89	7.7	–	···
2	19890109	32					–	···
	···							

Table 1
An example of a temporal database. This simple database shows the characteristics of a medical temporal database.

among more than three attributes are difficult to study even if visualization interfaces are used.

2 *Irregularity of temporal intervals*. Temporal intervals are irregular. Although most of the patients will come to the hospital every two weeks or once a month, physicians may not make laboratory tests at each visit. When a patient has an acute fit or suffers from acute diseases, such as pneumonia, laboratory examinations will be made every one to three days. On the other hand, when the status is stable, these tests may not be done for a long time. Patient ID 1 is a typical example. Between 04/30 and 05/08/1986, he suffered from a pneumonia and was admitted to a hospital. Then, during the therapeutic procedure, laboratory tests were made every 2-4 days. On the other hand, when he was stable, such tests were ordered every one or two years.

3 *Missing values*. In addition to irregularity of temporal intervals, datasets have many missing values. Even though medical experts will make laboratory examinations, they may not take the same tests in each instant. Patient ID 1 in Table 1 is a typical example. On 05/06/1986, a medical physician selected a specific test to confirm his diagnosis. So, he will not choose other tests. On 01/09/1989, he focused only on GOT, not other tests. In this way, missing values will be observed very often in clinical situations.

These characteristics have already been discussed in the area of KDD [Fayyad, 1996]. However, in real-world domains, especially domains in which follow-up

studies are crucial such as medical domains, these ill-posed situations will be distinguished. If one wants to describe each patient (record) as one row, then each row will have many attributes, depending on how many times laboratory examinations are made for each patient. It is notable that although the above discussions are based on medical situations, similar situations may occur in other domains with long-term follow-up studies.

Coding systems

Another difficulty is that medical databases will include specific codes. Medical informatics has a long history for coding systems, which originally comes from statistics of health care. Several important coding systems have been established and are widely used in HIS (Table 2) [Bemmel, 1997].

Table 2
Coding systems.

Name	Contents
ICD-9, 10	International classification of diseases
ICPC	International classification of primary care
DSM	Diagnostic and statistical manual for mental disorders
SNOMED	Systematized nomenclature of human and veterinary medicine
ICD-O	International classification of diseases for oncology
CPT	Current procedural terminology
ICPM	International classification of procedures in medicine
RCC	Read clinical classification
ATC	Anatomic therapeutic chemical code
MeSH	Medical subject headings
DRG	Diagnosis related groups

One of the major coding systems is the ICD (International Classification of Diseases), which represents each disease as a combination of letters and numbers. Table 3 shows an example of ICD codes for neurological diseases. It is easy to see that these codes are difficult to interpret and should be preprocessed.

Table 3
Examples of ICD-10 code.

Code	Definition
G30	Alzheimer disease
G30.0	Alzheimer disease, early onset
G30.1	Alzheimer disease, late onset
G30.8	Alzheimer disease, others
G30.9	Alzheimer disease, unspecified

For preprocessing, data miners should recognize that most coding systems have their own objectives. While ICD focuses on hierarchical structure of etiology,

SNOMED focuses on multiple aspects and has 11 axes to form a complete hierarchical classification system (Table 4) [Snomed, 1994]. Preprocessing may be dependent on the applied coding system.

Table 4
Eleven axes of SNOMED International.

Axis	Definition
T	Topography
M	Morphology
L	Living organisms
C	Chemicals
F	Function
J	Occupation
D	Diagnosis
P	Procedure
A	Physical agents, forces, activities
S	Social context
G	General

Most HIS databases use such coding systems and thus, raw data retrieved from such systems includes codes that are difficult to understand. Since transformation procedures depend on the coding systems, data miners should always note what coding systems have been used and they should transform these codes into categorical attributes correctly. Usually, codes are used to describe decision attributes, so extreme care is advised.

Translating medical problems into special forms of knowledge

Medical experts will view data mining as hypothesis generation and they need simple hypotheses. For this purpose, simple rule induction such as association rules, are applicable. However, a large database will generate a huge number of rules, most of which correspond to common sense. Thus, more sophisticated procedures may be required, such as coupling of rule induction with ontological reasoning and introduction of interestingness or surprisingness. Another important issue is that medical concepts and decisions are hierarchical. Although several researchers focus on this hierarchical nature [Tsumoto, 1998a], generalized techniques have not been derived yet.

Used data mining technique and results

Medical applications of data mining are a growing area in medical informatics. However, most of the studies, especially machine learning applications, fall into induction of predictive classification rules [Cooper, 1992; Lavrač, 1997;Zupan, 1999]. Although several interesting results can be obtained by such machine learning applications and rough set approaches [Polkowski, 1998], they mainly

focus on generation of rules with high predictive power, which can be called rules of general information. Unfortunately, those rules which have general knowledge are often corresponding to common sense in domain knowledge. It should be strongly pointed out that results from knowledge discovery processes do not always overlap with such rules of common sense knowledge, but rather with specific rules, which include new, interesting and unexpected patterns. In this section, we focus on three methods, the results of which suggest the future directions of medical data mining. For other studies, see [Cooper, 1992; Lavrač, 1997;Zupan, 1999].

Complications of cardioangiography

[Harris, 1984] applied exploratory data analysis [Tukey, 1977], statistical methods and a tree induction method to a database including 995 cases of cardioangiography, which were collected from five army hospitals. This dataset is described by 14 attributes (9: categorical, 5: continuous), which can be classified as a r-database.

Tree induction is a kind of extension of conventional decision-tree induction [Breiman, 1984], called sequential multiple regression analysis, where each node in the tree represents a multiple logistic regression function. In 995 cases, 42 cases included complications with angiography, and Harris examined important factors for complications. From statistical analysis (t-test and chi-square test), the name of the hospital, the gender (female), the location of inserted catheter and the period for angiography are significant factors for prediction of the complications. From the regression analysis, the gender (female), the period for angiography, the institute and the past history for myocardial infarction are significant factors. Finally, the decision tree selects gender as the first attribute. For male and female, the second factor is the name of the hospital and the location of catheter, respectively.

Harris argues that the most important discovery of this analysis is that the gender is selected as one of the most important factors. From this analysis, he reached one important hypothesis related with the gender: one of the main factors may be body size, which is related to the radius of arteries. If the radius of arteries is small, the probability of complications may be high. Since such body size is not included in a database, the gender is selected as an important factor with the association with body size.

This work suggests several important directions of medical data mining:

- Unexpected attributes will lead to the discovery of a new hypothesis.
- Detection of unexpected patterns and their interpretation needs deep background knowledge.
- Databases collected from different institutes will generate different patterns from each databases and some conflicts will be resolved.

Analysis of Meningoencephalitis

[Tsumoto, 1995; Tsumoto, 1999] applied a rule induction method based on variable precision rough models [Ziarko, 1993], decision trees [Breiman, 1984], and statistical methods, including multivariate analysis [Anderson, 1984;Andersen, 1991] to a database on meningoencephalitis, which is collected by Tsumoto as an r-database. All the induced results were interpreted by the author, a neurologist, which showed us that rule induction methods generated rules which are unexpected but interesting to medical experts, whereas decision tree methods and statistical methods acquired knowledge, which matches medical experts' knowledge.

The common datasets collect the data of patients who suffered from meningitis and were admitted to the department of emergency and neurology in several hospitals. These data are collected from the past patient records (1979 to 1989) and the cases in which the author made a diagnosis (1990 to 1993). The database consists of 121 cases and all the data are described by 38 attributes, including present and past history, laboratory examinations, final diagnosis, therapy, clinical courses and final status after the therapy.

A rule induction method based on the variable precision rough set model generated 67 rules for viral meningitis and 95 rules for bacterial meningitis, which included the following rules unexpected by domain experts as shown in Table 5. Especially, rules from 5 to 7 are new induced results, compared with those discovered by Tsumoto and Ziarko [Tsumoto, 1995]. In this study, grouping of attributes was applied. The attribute 'risk factor' has 12 values and these attribute-value pairs did not contribute to rule generation. Thus, this attribute was transformed into the binary attributes. After this transformation, these rules were obtained, which suggests that such a transformation is important to induce rules interesting to domain experts.

Table 5
Induced rules for meningitis.

1	[WBC < 12000]∧[Sex=Female]∧[CSF_CELL < 1000]	⇢ Viral	(Accuracy:0.97, Coverage:0.55)
2	[Age ≥ 40]∧[WBC ≥ 8000]	⇢ Bacterial	(Accuracy:0.80, Coverage:0.58)
3	[WBC ≥ 8000]∧[Sex=Male]	⇢ Bacterial	(Accuracy:0.78, Coverage:0.58)
4	[Sex=Male]∧[CSF_CELL ≥ 1000]	⇢ Bacterial	(Accuracy:0.77, Coverage:0.73)
5	[Risk_Factor=n]	⇢ Viral	(Accuracy:0.78, Coverage:0.96)
6	[Risk_Factor=n]∧ [Age <40]	⇢ Viral	(Accuracy:0.84, Coverage:0.65)
7	[Risk_Factor=n]∧ [Sex=Female]	⇢ Viral	(Accuracy:0.94, Coverage:0.60)

These results show that sex, age and risk factor are very important for diagnosis, which has not been examined fully in literature [Adams, 1993]. Thus, according to these results, the author re-examined relations between sex, age, risk

factor and diagnosis and discovered the following interesting relations between them[3]:

1 The number of examples satisfying [Sex=Male] is equal to 63, and 16 of 63 cases have a risk factor: 3 cases of DM (Diabetes Melitus), 3 cases of LC (Liver Cirrohsis) and 7 cases of sinusitis.
2 The number of examples satisfying [Age≥40] is equal to 41, and 12 of 41 cases have a risk factor: 4 cases of DM, 2 cases of LC and 4 cases of sinusitis.

It is notable that these results are also statistically significant by using statistical tests [Tsumoto, 1999]. DM an LC are well-known diseases in which the immune function of patients will become very low. Also, sinusitis has been pointed out to be a risk factor for bacterial meningitis [Adams, 1993]. It is also notable that males suffer from DM and LC more than females.
In this way, re-examination of databases according to the induced rules discovered new knowledge about meningitis. This empirical study also suggests that patterns unexpected but interesting to domain experts lead to a new discovery and that interpretation of such patterns needs deep background knowledge.

Association rules

[Brossette, 1998] applied association rule discovery to a database on hospital infection control and public health surveillance, which are essentially temporal databases. They state one of the problems with application of association rules: high-support and high-confidence rules will be less useful than high-support, low-confidence rules, which may be new, unexpected, and interesting patterns and they propose the following process for analyzing surveillance data:

– For each time-slice or partition of data, discover all high-support association rules.
– For each rule discovered in the current partition, compare the confidence of the rule from the current partition to the confidences of the rule in previous partitions.
– If the confidence of the rule has increased significantly from a previous partition, or previous partitions to the current partition, report this finding as a significant event.

Experimental results were obtained to analyze *Psudomonas aeruginosa* infection control data collected over one year (1996) at the University of Alabama at Birmingham Hospital. Experiments using one-, three-, and six-month time partitions yielded 34, 57 and 28 statistical significant events, respectively. Although not all of these events are clinically significant, several patterns show potentially significant shifts in the occurrence of infection or antimicrobial resistant patterns of *P. aeruginosa*.

3 These results have not been reported in the literature on neurology. Thus, although these discoveries may not be general knowledge, at least they include knowledge specific to this dataset. It can be called discovery of knowledge dependent on the context of data collection.

Their results are interesting not only in a medical context, but also in association rule discovery. The authors reported that for each time partition, more than 2,000, 12,000 and 20,000 rules were obtained. However, from the relations between rules of different partitions, only 34, 57 and 28 events were interesting patterns. This work suggests the following two points:

– Most of the induced results are not unexpected and not interesting.
– Interesting patterns may be obtained by comparison between rules in different setting.

Action taken by data owners

Three empirical studies show that medical experts try to interpret unexpected patterns with their domain knowledge, which can be viewed as hypothesis generation. In [Harris, 1984], gender is an attribute unexpected by experts, which led to a new hypothesis that body size will be closely related with complications of angiography. In [Tsumoto, 1995;Tsumoto, 1999], gender and age are unexpected attributes, which triggered reexamination of datasets and generated a hypothesis that immunological factors will be closely related with meningitis.

These actions will be summarized into the following three patterns:

1 If induced patterns are completely equivalent to domain knowledge, then the patterns are common sense.
2 If induced patterns partially overlap with domain knowledge, then the patterns may include unexpected or interesting subpatterns.
3 If induced patterns are completely different from domain knowledge, then the patterns may be meaningless.

Then, the next step will be the validation of a generated hypothesis: a dataset will be collected under the hypothesis in a prospective way. After the data collection, statistical analysis will be applied to detect the significance of this hypothesis. If the hypothesis is confirmed with statistical significance, these results will be reported.

Another action is to develop a decision support system with generated rules. There are very few studies on this topic: [Tsumoto, 1998b] reported one preliminary work on the relations between knowledge discovery and decision support. (See also Section 2.3.2, Decision support for medical diagnosis).

Future work

It is notable that data mining tasks achieved in the medical domain are still limited. Most studies, including machine learning applications, deal with simple tables, which are small pieces of information included in patient data of HIS. In those systems, a large amount of temporal data is left unanalyzed. Even

statistical techniques cannot deal with large quantities of temporal data, which will be a good opportunity to develop a new method from the data mining side.

Although in HIS mainly numerical data are stored, all other information, such as physical examinations and medical images will be included as electronic patient records [Bemmel, 1997]. Thus, discovery of patterns from those data and images will become an important research area of medical data mining. (see Sections. 5.5.3, Image mining and 5.5.4, Video mining).

Furthermore, the standardization of HIS established by ISO TC215 will enable us to introduce knowledge discovery in databases from different hospitals as distributed data mining. In medicine we have a huge amount of data in heterogeneous environments, which are waiting for further sophisticated techniques.
In summary, the following future research directions may be important for medical data mining techniques to be widely accepted in medicine.

- Discovery of relations between rules induced from databases.
- Coupling of medical knowledge with rules induced from databases.
- Discovery of temporal diagnostic patterns in HIS.
- Discovery of short-term effects and long-term effects of therapeutic procedures in HIS.
- Rule discovery in databases collected from different hospitals.
- Discovery of new diseases.

Although these tasks are very difficult to achieve, existing techniques are not applicable and new techniques should be introduced.

In the 21st century, medical data mining and knowledge discovery will be a hot research topic in medical informatics just after all the techniques for electronic patient records have been established. Problems with medical data mining may inspire the development of general data mining techniques. Then, the growth of data mining techniques will support this hot research topic in medical informatics, which will finally contribute to the development of medical research and practice.

References

- Abiteboul, S., R. Hull, V. Vianu. (1995). Foundations of Databases. Addison-Wesley, New York
- Adams, R.D., M. Victor. (1993). Principles of Neurology. 5th edition. McGraw-Hill
- Agrawal, R., T. Imielinski, A. Swami. (1993). Mining Association Rules between Sets of Items in Large Databases. Proceedings of the 1993 International Conference on Management of Data (SIGMOD 93). pp207-216

- Altman, D. (1991). Practical Statistics for Medical Research. Chapman and Hall
- Andersen, E.B. (1991). The Statistical Analysis of Categorical Data. Second, Revised and Enlarged Edition. Springer Verlag
- Anderson, T.W. (1984). An Introduction to Multivariate Statistical Analysis. John Wiley & Sons, New York
- Bemmel, J. van, M.A. Musen. (1997). Handbook of Medical Informatics. Springer Verlag, New York
- Breiman, L., J. Freidman, R. Olshen, C. Stone. (1984). Classification And Regression Trees. Wadsworth International Group
- Brossette, S.E., A.P. Spragce, J.M. Hardin, K.B. Wates, W.T. Jones, S.A. Moser. (1998). Association Rules and Data Mining in Hospital Infection Control and Public Health Surveillance. Journal of American Medical Informatics Association **5**:375-381
- College of American Pathologist. (1994). SNOMED. College of American Pathologist, Chicago
- Cooper, G.F., C.F. Aliferis, J. Aronis, B.G. Buchanan, R. Caruana, M.J. Fine, C. Glymour, G. Gordon, B.H. Hanusa, J.E. Janosky, C. Meek, T. Mitchell, T. Richardson, P. Spirtes. (1992). An Evaluation of Machine-Learning Methods for Predicting Pneumonia Mortality. Artificial Intelligence in Medicine **9**:107-139
- Fauci, A.S., E. Braunwald, K.J. Isselbacher, J.B. Martin. (eds.). (1997). Harrison's Principles of Internal Medicine, 14th Edition. McGraw Hill, New York
- Fayyad, U.M. et al. (eds.). (1996). Advances in Knowledge Discovery and Data Mining. AAAI Press
- Harris, J.M. (1984). Coronary Angiography and its Complications - The Search for Risk Factors, Archives of Internal Medicine **144**:337-341
- Kleinbaum, D.G., L.L. Kupper. (eds.). (1982). Epidemiologic Research: Principles and Quantitative Methods. John Wiley & Sons, New York
- Lavrač, N., E.T. Keravnou, B. Zupan. (eds.). (1997). Intelligent Data Analysis in Medicine and Pharmacology. Kluwer, Dordrecht
- Levinson, W.E., E. Jawetz. (1996) Medical Microbiology & Immunology: Examination and Board Review. 4th Edition. Appleton & Lange
- Lucus, P.J.F., A. Abu-Hanna. (eds.). (1999). Special Issue on Prognostic Models in Medicine. Artificial Intelligence in Medicine **15** (2)
- Piatetsky-Shapiro, G., W. Frawley. (1991). Knowledge Discovery in Databases. AAAI Press
- Plum, F. (1992). The Diagnosis of Stupor and Coma. 3rd Edition. F.A. Davis and Co.
- Polkowski, L., A. Skowron. (1998). Rough Sets in Knowledge Discovery. Vol. 1 and 2. Physica Verlag, Berlin

- Rothman, K.J., S. Greenland. (eds.). (1998). Modern Epidemiology. Lippincott-Raven Publishers
- Tsumoto, S., W. Ziarko, N. Shan, H. Tanaka. (1995). Knowledge Discovery in Clinical Databases Based on Variable Precision Rough Set Model. Proceedings of the Eighteenth Annual Symposium on Computer Applications in Medical Care, Journal of the American Medical Informatics Associations **2**, supplement. pp270-274
- Tsumoto, S. (1998a). Extraction of Experts' Decision Rules from Clinical Databases using Rough Set Model. Journal of Intelligent Data Analysis **2** (3)
- Tsumoto, S. (1998b). Automated Knowledge Acquisition based on Rough Sets and Attribute-Oriented Generalization. Proceedings of the AMIA Fall Symposium. Journal of AMIA **5**. Supplement
- Tsumoto, S. (1999). Knowledge Discovery in Clinical Databases - An Experiment with Rule Induction and Statistics. In: Z. Ras. (ed.). Proceedings of the Eleventh International Symposium on Methodologies for Intelligent Systems (ISMIS'99). Springer Verlag (in press)
- Tukey, J. (1977). Exploratory Data Analysis. Addison Wesley, New York
- Westfall, P.H., S. Stanley Young. (1993). Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment. John Wiley & Sons, New York
- Ziarko, W. (1993). Variable Precision Rough Set Model. Journal of Computer and System Sciences **46**:39-59
- Zupan, B., N. Lavrač, .E. Keravnou. (1999). Special Issue on Data Mining Techniques and Applications in Medicine. Artificial Intelligence in Medicine, **16** (1)

2.3.2 Decision support for medical diagnosis

Bert Kappen[1], *Wim Wiegerinck*[2], *Edith ter Braak*[3]

Introduction

'Do we need computerized diagnostic decision support systems in medical practice?'

Problems in modern medicine are often very complex, and evidence for the best choice to be made is often lacking. Decisions made by physicians can vary (for one physician over time and between physicians), and are sometimes lacking explanation or 'rationalization' [Eddy, 1990; Berg, 1997]. Clinical examples of this phenomenon in making diagnoses are abundant. The body of potentially useful knowledge that is relevant to even a relatively narrow diagnostic area may be too large to justify an optimal (diagnostic) decision on the spot. Ironically, modern information technology (especially through the Internet) further increases the amount of available knowledge, potentially even further complicating this situation. Moreover, individual patients need 'individualized' decisions, because their characteristics differ from the 'average' and because of their individual wishes [Lilford, 1998]. Apparently, individualizing the general results of research may be cumbersome and time-consuming, while on the other hand, modern medical practice demands efficiency, cost-effectiveness and high technical quality. The derivation of diagnostic protocols is a main problem in health care. In some environments diagnostic support asset out in this proposal is unlikely to influence the physician's decisions, e.g. on a neurological intensive care unit, since the diagnosis is often obvious [Brigl, 1998]. In contrast, general internal medicine covers an enormous range of, sometimes relatively rare, diagnostic categories, hence the tendency of medicine to be divided in super specialties. A diagnostic decision support system covering general internal medicine may be appreciated by both generalists and super-specialists alike: by the generalist because this field of work typically covers a very broad range of diagnoses, by the super-specialist because he or she may not feel completely at ease outside the specific field of expertise. It will be readily understood that the above comprises an enormous task and challenge for modern medicine in general and individual doctors in particular, illustrating the need for decision support techniques. Obviously, computerized decision aids may be very promising from a theoretical point of view. However, the currently available systems have not yet been very successful and certainly their use is still not widespread and not established in daily routine. A variety of factors may be responsible for this:

– *Lack of accuracy*: Those current systems that intend to cover a broad diagnostic domain of medicine generally lack diagnostic accuracy [Berner, 1994; Berner, 1996]. This is mainly due to the levels of detail (e.g. diagnostic

1 Dr H.J. Kappen, bert@mbfys.kun.nl, Nijmegen University, Department of Medical Physics and Biophysics, Nijmegen, The Netherlands.

2 Dr W.A.J.J. Wiegerinck, wimw@mbfys.kun.nl, Department of Medical Physics and Biophysics, Nijmegen University, Nijmegen, The Netherlands.

3 Dr E.W.M.T. ter Braak, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

categories at the level of ICD-10) and completeness in the knowledge base [Shwe,1991; WHO, 1992]. In contrast, systems that are based on detailed modeling of knowledge, resulting in good performance, are restricted to a relatively narrow field [Heckerman, 1992a; Heckerman, 1992b].

- *Lack of transparency*: In the era of evidence based medicine the advice of 'a machine', functioning as a black box is unacceptable: an advice must be accounted for on the basis of research published in the peer-reviewed literature. The majority of conventional protocols and consensus guidelines also often fail to refer explicitly to the literature. Therefore, (diagnostic) advices suggested by a computerized tool should come with the appropriate references from the literature.
- *Users attitude*: In a subset of (potential) users there may be a misunderstanding about what computers can and cannot do for them. Generally, decision support systems need intelligent and responsible users, who are able to interpret the advice given and estimate its merit [Dijkstra, 1998]. This, however is not exclusively a matter of user's attitudes. Producers of decision support tools should take this issue into account as well, especially when designing the user interface and deciding which facilities are needed.
- *Lack of integration of information*: Patient oriented decision support needs data from several sources. A decision support system will generate new information (e.g. a diagnostic advise) through inference, using patient-specific information. Integration of information, multiple usability of patient data, integration of databases and knowledge bases are common problems when using a heterogeneous Hospital Information System (HIS). In practice, the completeness of patient information, and the accuracy and level of detail of diagnoses stored in the HIS is often very poor [Wiegerinck, 1997].
- *Lack of a controlled terminology*: This is a problem that might not even be solved completely in the near future. Most standard classification systems are at a general level, thus lacking the required detail, or specialized and therefore too limited to meet the needs for a broad decision support system [WHO, 1992; Boersma, 1993; WHO, 1992]. Furthermore, a standard classification is not always available, for instance for specific terminology used in text books.
- *Careful introduction*: Introduction of a decision support system should be done as carefully and thoroughly as it is for drugs that are new on the market. Oddly enough this tradition of careful introduction (and marketing !) is common in the field of therapeutics, but not quite as established for support tools in general and for diagnostics in particular. After introduction, the decision support system will need constant monitoring of users needs and maintenance to keep up with the latest results of medical research.
- *The need of an integrated clinical workstation*: The appropriate infrastructure and workstations are not yet available in all hospitals. Physicians will

need on-line support during the implementation of the various functionalities of a reliable clinical workstation, which integrates all the required information.

In conclusion, modern medicine is in need of computerized decision aids both to meet its own high standards and to keep pace with the stage of development in other domains such as manufacturing or the services industry. Although decision support appears to be exceptionally suitable for the medical domain, computer aided decision-making in medicine is still in its infancy. The development, implementation, assessment and further improvement of decision support systems in medicine still need a lot of research.

Our approach

Diagnostic reasoning in the medical domain is a typical example of reasoning with uncertainty. This uncertainty has different sources: missing patient information, uncertainty in medical tests results or observations, and the uncertainty about the physiological processes involved. A model on which a DSS is based should be able to deal with these uncertainties. The different systems that have been developed so far use a variety of modeling approaches which can be roughly divided into two categories, large rule-based systems and smaller probabilistic systems.

The large systems that attempt to cover the whole of internal medicine use a rule-based approach with some rather heuristic method to quantify uncertainty. These methods perform poorly in practice [Berner, 1994; Berner, 1996]. The main reasons are that the modeling of the relations between diseases and findings is at a very coarse level. Therefore, the diagnoses suggested by these systems are too superficial for clinical use. Secondly, the diagnostic process requires reasoning from causes to effects (diseases —› finding) and vice versa at the same time. The rule-based approach, together with the heuristics for uncertainty, is not well suited for such bi-directional reasoning.

For smaller systems, the probabilistic approach is typically used. The probabilistic approach has the important advantage of mathematical consistency and correctness. In particular Bayesian networks (see Section 6.2.10) provide a powerful and conceptual transparent formalism for probabilistic modeling (see e.g. [Lauritzen, 1988; Pearl, 1988; Castillo, 1997]). In addition, they allow for easy integration of domain knowledge and learning from data. Systems that are based on detailed modeling have been restricted to a relatively small domain [Heckerman, 1992a; Heckerman, 1992b]. The reason for this is that Bayesian networks will become intractable for exact computation if a large medical domain is modeled in detail. Intractability means that the time required for computation scales exponentially fast with the number of variables in the model.

To proceed one has to rely on approximate computations. Recently, variational methods for approximation are becoming increasingly popular [Saul, 1996; Jaakkola, 1997; Barber, 1999]. An advantage of variational methods techniques is that they provide bounds on the quantity of interest in contrast to stochastic sampling methods which may yield unreliable results due to finite sampling times. Until now, variational approximations have been less widely applied than Monte Carlo [4] methods, arguably since their use is not so straightforward. We argue that variational methods are indeed applicable to large, detailed Bayesian networks for medical diagnosis constructed by human experts. Although the formalism of Bayesian networks is very powerful, the construction of networks for medical diagnosis is not straightforward. A learning approach depends crucially on the availability of high quality patient data. In particular, rare disorders should be well covered. In general, unfortunately, this is rather the exception than the rule [Wiegerinck, 1997]. Therefore, to reach a successful diagnostic decision, a support system requires explicit modeling effort by human experts. The existing medical literature is not sufficient to define the probabilistic model; not all probabilistic relations between variables have been documented. However, medical literature provides a useful starting point for model design. Once a minimal performance has been thus obtained, the model can be improved by learning from patient data.

Probabilistic modeling in the medical domain

We will outline here what the structure of a broad and detailed Bayesian network will typically look like. This is based on an extrapolation of our current modeling experiences. Details of the medical domain are beyond the scope of this paper and are discussed elsewhere [Burg, 1999]. The variables to consider in the network are of different types. There are disease variables, which are typically of the binary type, signaling whether a disease is present or not. The findings encode the results from laboratory measurements, physical examination etc. As a simplification, these variables are discretized, with medically relevant cut-off points. In practice, such discretization does not lead to significant loss of information. In addition, there are prior variables that describe the patient, such as sex and age. In constructing the graph for the Bayesian network, human experts mostly use 'causal' relationships between variables as a guideline (the arrows in Figure 1). Often, the expert can relate (large numbers of) variables via additional hidden variables. These hidden variables may represent pathophysiological variables that are known to have certain relations to the observable variables, but are themselves not accessible during clinical investigation. Often, the use of hidden variables results in a simplified and more transparent network. The majority of probabilistic relations between the variables involve only a small number of parents. Consequently, modeling using explicit probability tables is feasible. These are estimated on the basis of data in the literature or on

4 Method making use of random numbers and probability statistics. See Sobol, I.M. (1994). A Primer for the Monte Carlo Method. Inst. Mathematical Modeling, Moscow.

'educated guesses' based on local statistics and/or experience if no data from the literature are available. Medical experts tend to divide knowledge concerning a medical domain into subdomains with a relatively small overlap. Therefore, the network will typically have a modular structure (cf. Figure 1). Each module represents a disease with its relevant findings. In practice, the modules are rather small, containing between 20 and 50 variables. Different modules are connected via shared variables (e.g. pathophysiological variables that are relevant in different modules), common prior nodes, and or common findings nodes. The computational complexity of the network N1 consisting of the modules and their parents (black nodes in Figure 1) is tractable, i.e. will not require excessive computation time.

The probabilistic relations for the findings require somewhat more care. For example, 'hemoglobin level' (Hb) is a variable whose value is affected by many diseases. Such nodes may have parents in many subdomains. This makes the use of a conditional probability table infeasible, as the size of the table grows exponentially in the number of parents. Fortunately, this is neither necessary, since medical experts are likely to agree with a 'sum of univariate relations' between this finding and its parents. Such simplified conditional probability tables require only the specification of order k parameters, where k is the number of parents. Even though the conditional probability tables are modeled in a compact way, inference is still intractable.

Variational approximations

In general, the problem of inference is to find the conditional probability of some variables (diseases, other findings) given some findings. For instance, given the measurement of high blood pressure in a patient, one may want to compute the probabilities of the different possible causes (diseases) for this observation. As mentioned before, computation of most realistic models cannot be done exactly in a reasonable time. Therefore one must invoke approximations.

The model is a high dimensional probability distribution $P(s1,..,sn)$ over a set of variables. In the variational method, this distribution is approximated by a tractable distribution $Q(s1,..,sn)$. The model Q is simpler than the model P and computing with Q is fast. The task is to find the distribution Q that best approximates our original model P. One can define an error criterion, that quantifies how different the distributions P and Q are. The best Q is found by minimization of this error with respect to the parameters that describe Q. This optimization requires the simultaneous solution of a set of n coupled non-linear equations, where n is the number of variables in the model. Its solution can be obtained efficiently.

The quality of the approximation depends strongly on the structure of Q, meaning which variables in the model Q are correlated. The simplest approach is the

so called mean-field approach, in which all variables in Q are independent. This assumption yields the fastest solution, but with the largest error. The other extreme is to take for Q a graphical structure that is identical to P. In that case one obtains the solution $Q=P$, i.e. error is zero, as expected. But the computational time for this solution is as bad as computing with the model P directly. This solution is only of theoretical interest, but it indicates that the variational approach using structure interpolates between the standard mean field theory and the exact solution. In general one must choose a structure for Q that is a good compromise between approximation error and computational efforts.

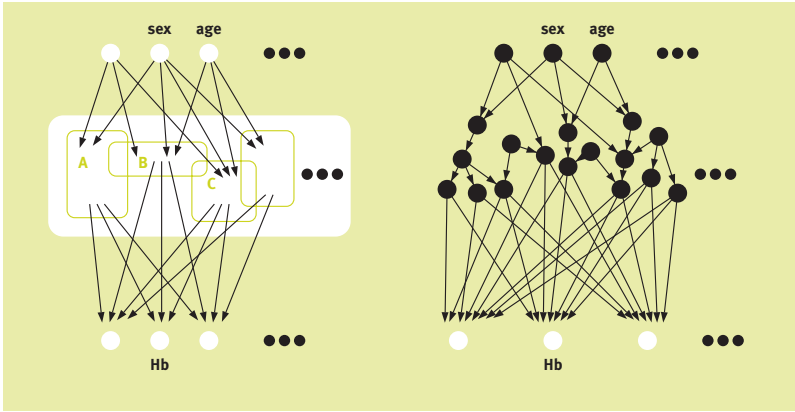


Figure 1
Modular and graphical network structure. Left: modular structure of the network. A, B, C. represent (overlapping) subdomains. Each subdomain is modeled by a number of nodes (cf. right figure) representing variables that are relevant in that domain. The upper nodes, e.g. 'sex' and 'age' represent common ancestors of nodes in several subdomains. The lower nodes, e.g. 'Hb' represent common children of nodes in several subdomains (e.g. related to anaemia). Right: underlying graphical structure of same network. Filled circles: nodes in subdomains and their common ancestors. Open circles: common children.

In Figure 2 we plotted the maximal error as a function of the network size for an artificially generated network. We also plotted the required computer time for exact and approximate inference as a function of the network size. We conclude that variational methods using structure significantly improves the quality of approximation, within feasible computer time. In a network with tractable substructures, as can be expected in medical diagnosis, these substructures provide a useful starting point for the approximating model. For more information on the variational approach see [Kappen, 2001; Wiegerink, 1999].

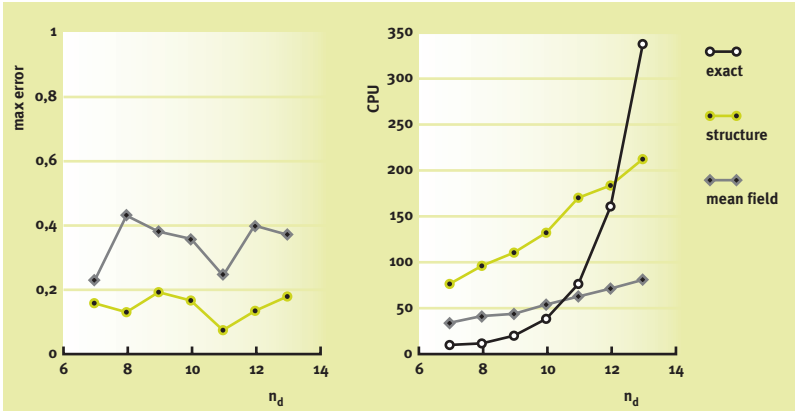


Figure 2
Left: The maximal error as a function of the network size. Right: CPU-time in Matlab seconds for exact and approximate inference as a function of the network size.

Promedas, a demonstration of DSS

Promedas (PRObabilistic MEdical Diagnostic Advisory System) is a decision support system that we are developing for the problem of anaemia. The aim of Promedas is to assess the possibility of using Bayesian networks for large medical domains and to assess the usefulness of approximate methods for diagnostic decision support.

The problem domain anaemia is chosen because we expect that the computational problems described in the previous sections will be encountered in this domain. For instance, anaemia can be subdivided in a large number of subdomains, each of which share a large number of findings. Furthermore, anaemia is a common medical problem. This facilitates evaluation in practice. To cover the domain completely, we expect that approximately 1,000 nodes will be needed. To develop Promedas, we use our internally developed software environment, called BayesBuilder. BayesBuilder has graphical tools for network construction, evaluation, and maintenance. So far, Promedas covers megaloblastic anaemia. It is currently based on a network of 91 variables, and is still tractable for exact algorithms. Promedas consists of a graphical user interface (GUI) to enter patient data and for diagnostic consultation (see Figure 3). It provides a differential diagnosis, i.e. the probabilities of potentially relevant diagnoses and the probabilities of potentially involved underlying mechanisms (e.g. pathophysiology) as percentages (ranked in descending order). These probabilities are computed on the basis of the available findings entered in the system. In addition, Promedas computes which additional tests it expects to be most informative to decide about a diagnosis, specified by the user.

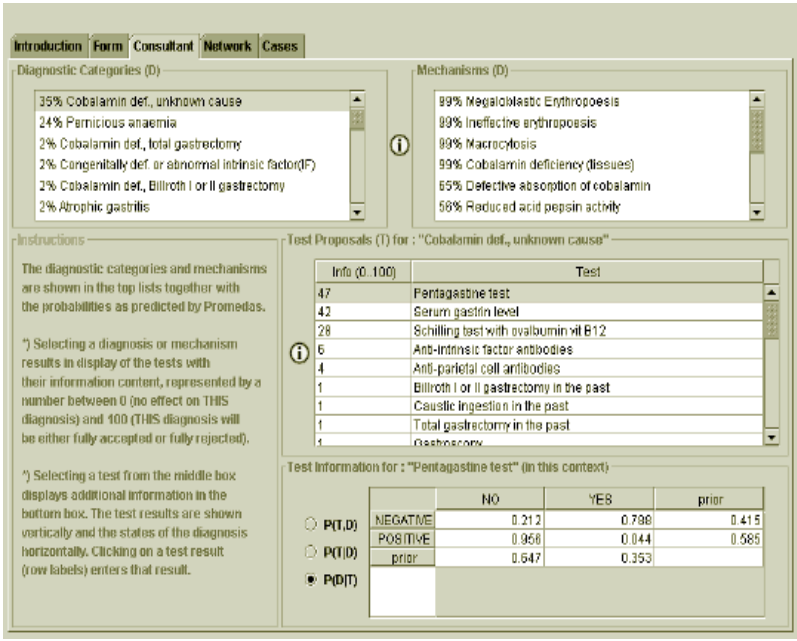


Figure 3
The Promedas diagnostic decision support system.

This information is computed given the P values of the variables previously entered and is defined as $I(D, T) = D,TP(D, T)\ \ln(P(D, T=)P(D)P(T))$ with $P(D, T)$ the joint probability of diagnosis and test result, and $P(D)$, $P(T)$ the marginal probabilities of diagnoses and tests, respectively.
These probabilities are computed by marginalizing over all the missing variables in the network. The information is normalized between 0 and 100, and displayed in descending order. In addition, Promedas provides help information, medical background information and pointers to the literature.

Utilization

Taking into account the need for decision support systems in general and diagnostic decision support in particular, we strongly believe that a diagnostic decision support system is viable and, eventually, marketable. However, even a pilot regarding the implementation and assessment of a diagnostic decision support system that covers only a relatively narrow diagnostic field (i.e. anaemia) will need careful, preferably step-wise, introduction to its target users, followed by continuing support and mutual feedback. It is expected that this will result in growing acceptance and enthusiasm by its target users and finally will lead to wide-spread use. The physicians who participate within this project work in daily hospital practice as well. They have good contacts in the medical community in many fields of medicine (general internal medicine, oncology, endocrinology, haematology) within Utrecht University Hospital, affiliated (regional) hospitals and various professional circles. In addition, the user group includes specialists in internal medicine from other academic hospitals. Therefore, we feel that we are able to 'market' and to follow up a diagnostic decision support system at least for research purposes (assessment) in The Netherlands. Wide spread acceptance of computer aided diagnostic decision support tools will probably need some 'trend setters', who are most likely to be found in the academic circle.

Discussion

The development of a DSS for comprehensive medical diagnosis in internal medicine represents a great challenge for AI. A broad and detailed probabilistic network is intractable for exact inference in this context. It is currently unknown, whether variational or other approximate methods are sufficiently powerful to provide a practical solution. The 'quality of approximation' is to a large extent a user defined (medical) issue, since 1. comparison with exact inference is not possible due to the size of the networks and 2. errors in the approximation will be judged as acceptable not just on their numerical values, but more importantly on their medical implications. The only way to assess the usefulness of approximate methods for modeling medical domains is by actually building such a system and evaluating it in use. The Promedas model must be extended

to 500-1000 variables in order to be able to address this issue properly. Further reading on graphical models can be found in [Pearl, 1988; Jensen 1996]. Further reading on diagnostic decision support systems can be found in [Brigl, 1998; Berg, 1997]. The Promedas project is supported by the Dutch Technology Foundation STW.

Future

Our experience with Promedas has convinced us that the use of Bayesian methods for medical diagnosis is superior to a rule-based approach. The main difference is that in the Bayesian approach the medical domain knowledge (what is the relation between symptoms and diseases) and the different type of operations that one may want to perform (for instance, to generate a list of possible diagnoses given some patient data) are separated. The modeling of the domain is where all the medical assumptions enter. The operations are purely mechanical (although non-trivial) and their outcome (for instance a differential diagnosis for a patient) follow logically from the model. No further assumptions are made. This implies that when the user does not agree with the outcome of the system, the cause is a modeling error and can be revised.

In the rule-based approach, no such distinction between model and operations exists. Instead, the medical knowledge enters in each of the rules. Since the number of rules that are required to cover all possible patient conditions is very large, the maintenance, assessment and revision of medical knowledge in rule-based systems, such as MYCIN and DXPLAIN is almost impossible.

Despite the clear advantage of the Bayesian approach, the construction of a 'correct' graphical model describing a large medical domain is very time-consuming. In addition, it is a task that can only be performed reliably by medical experts. In our project, three specialists in internal medicine participate. Each of them is specialized in a particular subdomain of internal medicine and will model that part. However, due to their clinical obligations they can only dedicate part of their time. Therefore, progress is necessarily limited by these practical constraints. However, we are quite convinced that despite these problems, we can model a large subdomain of internal medicine in the coming two years.

A crucial future issue is the electronic patient dossier. The absence of case based electronic dossiers has prevented us in the past from 'learning' the learn part of the model directly from patient data. Data will never replace explicit modeling, but it is clear that a combination of domain knowledge and data will give superior results. In our approach, we will build on the basis of domain knowledge only until the model is large enough to test with some users (other internists). At that stage we can also collect prospective data and use this to further test and refine our model.

We have developed a software tool, called BayesBuilder, that allows for efficient model building. We expect that the software will soon have matured sufficiently

to be used in parallel by a larger community of medical experts, thus accelerating the design process.
Promedas focuses on a detailed modeling of complex disease mechanisms in internal medicine. It is used for the education of medical students and will be used to assists internists during diagnosis. Other medical domain applications include (web based) medical diagnosis to be used by patients. In this case, a course model of the general practitioner's expertise could help patients to decide whether they should consult a doctor as well as which hospital provides the best expertise.

References

- Barber, D., W.A.J.J. Wiegerinck. (1999). Tractable Variational Structures for Approximating Graphical Models. In: M.S. Kearns, S.A. Solla, D.A. Cohn (eds.). Advances in Neural Information Processing Systems **11**. MIT Press
- Berg, M. (1997). Rationalizing Medical Work, Decision-Support Techniques and Medical Practices. MIT Press
- Berner, E.S., G.D. Webster, A.A. Shugerman, J.R. Jackson, J. Algina, A.L. Baker, E.V. Ball, C.G. Cobbs, V.W. Dennis, E.P. Frenkel, L.D. Hudson, E.L. Mancall, C.E. Racley, O.D. Taunton. (1994). Performance of four Computer-Based Diagnostic Systems. New England Journal of Medicine **330** (25):1792-1796
- Berner, E.S., J.R. Jackson, J. Algina. (1996). Relationships among Performance Scores of four Diagnostic Decision Support Systems. Journal of the American Medical Information Association **3** (3):208-15
- Boersma, J., R.S. Gebel, H. Lamberts. (1993). International Classification of Primary Care, Short Titles (translated into Dutch) en Dutch Subtitles. Nederlands Huisartsen Genootschap
- Brigl, B., P. Ringleb, T. Steiner, P. Knaup, W. Hacke, R. Haux. (1998). An Integrated Approach for a Knowledge-Based Clinical Workstation: Architecture and Experience. Methods of Information in Medicine **37**:16-25
- Burg, W.J. ter, et al. (1999). A Diagnostic Advice System Based on Pathophysiological Models of Diseases. Medical Informatics Europe, Ljubljana, Slovenia. Accepted
- Castillo, E., J.M. Gutierrez, A.S. (1997). Hadi. Expert Systems and Probabilistic Network Models. Springer Verlag
- Dijkstra, J.J. (1998). On the Use of Computerised Decision Aids. PhD Thesis. State University of Groningen
- Eddy, D.M. (1990). The Challenge. Journal of American Medical Association **263**:287-290
- Heckerman, D.E., E.J. Horvitz, B.N. Nathwani. (1992a). Towards Normative Expert Systems: Part I, the Pathfinder Project. Methods of Information in Medicine **31**:90-105

- Heckerman, D.E., B.N. Nathwani. (1992b). Towards Normative Expert Systems: Part II, Probability-Based Representations for Efficient Knowledge Acquisition and Inference. Methods of Information in Medicine **31**:106-116
- Jaakkola, T.S., M.I. Jordan. (1997). Variational Methods and the QMR-DT Database. MIT Computational Cognitive Science. Technical Report 9701, Massachusetts Institute of Technology
- Jensen, F.V. (1996). An Introduction to Bayesian Networks. UCL Press
- Kappen, H.J., W.A.J.J. Wiegerinck. (2001). Mean Field Theory for Graphical Models. In: D. Saad, M. Opper. (eds.). Advanced Mean Field Theory. MIT Press
- Lauritzen, S.L., D.J. Spiegelhalter. (1988). Local Computations with Probabilities on Graphical Structures and their Application to Expert Systems. Journal of the Royal Statistical Society B **50**:154-227
- Lilford, R.J., S.G. Pauker, D.A. Braunholz, J. Chard. (1998). Decision Analysis and the Implementation of Research Findings. British Medical Journal **317**:405-409
- Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers
- Saul, L.K., T. Jaakkola, M.I. Jordan. (1996). Mean Field Theory for Sigmoid Belief Networks. Journal of Artificial Intelligence Research **4**:61-76
- Shwe, M.A., B. Middleton, D.E. Heckerman, M. Henrion, E.J. Horvitz, H.P. Lehman, G.F. Cooper. (1991). Probabilistic Diagnosis Using a Reformulation of the Internist-1/ QMR Knowledge Base. Methods of Information in Medicine **30**:241-255
- Whittaker, J. (1990). Graphical Models in Applied Multivariate Statistics. Wiley, Chichester
- WHO. (1992). International Classification of Diseases, 10th Revision, Clinical Modification ICD-10
- WHO. International Classification of Diseases for Oncology
- Wiegerinck, W.A.J.J., W. ter Burg, E. ter Braak, P. van Dam, Y.L. O, J. Neijt, H.J. Kappen. (1997). Inference and Advisory System for Medical Diagnosis. Technical Report. SNN
- Wiegerinck, W.A.J.J., D. Barber. (1998). Mean Field Theory Based on Belief Networks for Approximate Inference. In: L. Niklasson, M. Bod'en, T. Ziemke. (eds.). ICANN'98 Proceedings of the 8th International Conference on Artificial Neural Networks. Springer Verlag
- Wiegerinck, W.A.J.J., D. Barber. (1998). Variational Belief Networks for Approximate Inference. In: H. La Poutré, J. van den Herik. (eds.). Proceedings of the Tenth Netherlands/Belgium Conference on Artificial Intelligence (NAIC'98). Amsterdam. CWI
- Wiegerink, W.A.J.J., H.J. Kappen, E.W.M.T. ter Braak, W.J.P.P. ter Burg, M.J. Nijman, Y.L. O, J.P. Neijt. (1999). Approximate Inference for Medical Diagnosis, Foundation for Neural Networks, University of Nijmegen

2.3.3 Bioinformatics

Lodewyk Wessels, Marcel Reinders[1]

Introduction and history

Rapid developments in sequencing technology and the introduction of micro array technology have caused (and are still causing) a biotechnological revolution. The sequencing of various genomes (including the human genome) in conjunction with micro arrays enables the measurement of the genetic activity of these organisms. The newest branch in these technological developments is research towards measuring protein presence and activity with chip technology [Pandey, 2000]. Briefly, the major advances of the miniaturizations are:

- only a very small sample is necessary to perform measurements (eventually, this will make measurements on a single cell possible).
- many measurements can be performed simultaneously (all genes and protein levels, for example).

For the first time in history a complete picture of the activities within a cell has become available. With these technologies, we thus gain insight into the activities within a cell and how they are governed, which eventually allows us some degree of control over these activities. These include, for example, increasing the growth in a yeast cell or to design a drug against a disease such as cancer.

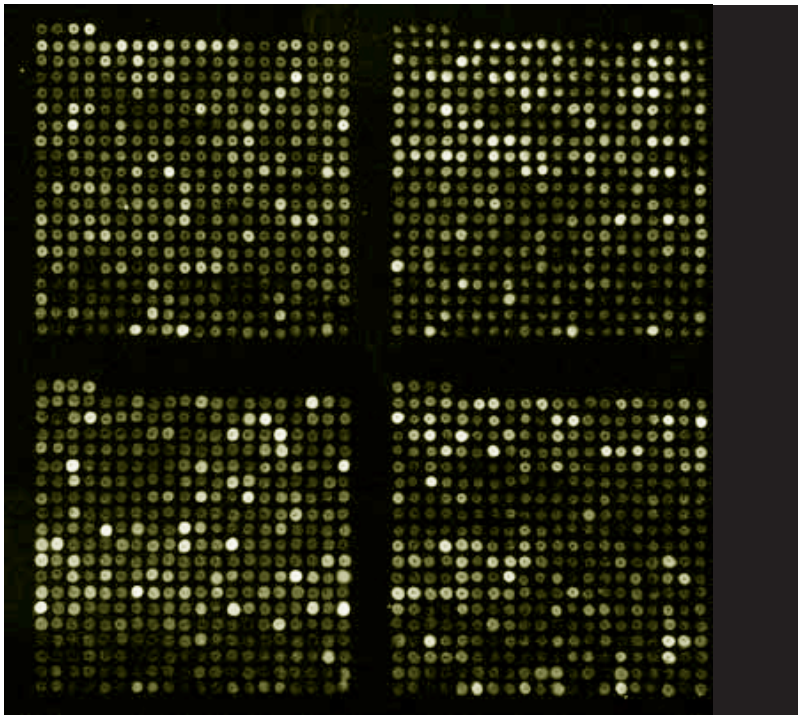


Figure 1
Micro array measurements of gene expression activities (mRNA activity).

1 Dr Ir L.F.A. Wessels
l.f.a.wessels@its.tudelft.nl,
Dr Ir M.J.T. Reinders,
m.j.t.reinders@its.tudelft.nl,
Information and Communication Theory Group, Department of Electrical Engineering, Faculty of Information Technology and Systems, Delft University of Technology, The Netherlands

Biomolecular scientists are extremely enthusiastic about these developments, but are now confronted with the availability of enormous amounts of data about a complex (many parameters, i.e. many genes) and heterogeneous (many elements, i.e. genes as well as proteins and metabolites) system. Traditional reductionist analysis, like pair-wise correlations, is insufficient to analyze this type of data. As a consequence, a new research field — called Bioinformatics — has emerged that is directed solely towards the generation of knowledge about these complex cellular systems from large amounts of molecular biological information[2].

Bioinformatics is an interdisciplinary research area that may be defined as the interface between biological and computational sciences. In general, the term 'bioinformatics' is not really well defined. One could say that this scientific field deals with the computational management of all kinds of biological information, whether it is about genes and their products, whole organisms or even ecological systems. Most of the bioinformatics work now being done can be described as analyzing biological data, although a growing number of projects deal with the organization of biological information. Most of the current Bioinformatics projects deal with structural and functional aspects of genes and proteins. Many of these projects are related to the Human Genome Project[3].

Clearly, bioinformatics has been strongly influenced by the micro array developments. Currently, many research teams all over the world actively produce high-throughput biomolecular data. Their data is collected and organized in databases specialized for particular subjects. Well-known examples are: GDB, SWISS-PROT, GenBank, and PDB. The latter — for example — deals with three-dimensional structures of biological molecules. Clearly, computational tools are needed to analyze the collected data in the most efficient manner. For example, many bioinformaticians are working on the prediction of the biological functions of genes and proteins (or parts of them) based on structural data.

Due to its interdisciplinary nature there are two types of Bioinformaticians: the 'miners' and the 'engineers'. The miners have biological training and know how to extract biological information from sequenced genomes. The engineers have a strong computational background and design the mining tools [Meyers, 2000].

Bioinformatics, the discipline

Basically, biomolecular science is the discipline that deals with unraveling the players within the central dogma for a specific organism. The central dogma describes the basic mechanisms that control the processes within a cell. Simply put, the central dogma states that the DNA encodes functional elements, the genes, that are activated by proteomic interactions. Upon activation the gene is

2 http://bioinformatics.weizmann.ac.il/cards/bioinfo_intro.html

3 http://www.nhgri.nih.gov

transcribed and translated into nucleotides that after modification and folding become proteins. These proteins interact with each other or catalyze metabolic reactions that take place within the cell. This results in the change of protein levels that finally affects the activity of the genetic information.
The research themes within bioinformatics are closely related to the different aspects within the central dogma. On an abstract level, we make the following distinction:

1 sequencing;
2 structure prediction;
3 function prediction from sequence data;
4 function prediction from gene activity levels.

Sequencing

Sequencing is the term that refers to the mapping of a genome of an organism, i.e. the determination of the DNA sequence (C, T, A, G's). The larger part of this technology is biochemical in nature. Since current techniques are only capable of sequencing relatively small parts of the DNA strain, it is cut into small overlapping parts. Since the cut is done at unknown positions (the sequence is not known), putting them back in the right order is a combinatorial problem.
Lots of computer power and consequent computational effort are being put into this combinatorial problem (the larger part of Celera, Craig Venter's sequencing company, consists of high performance computers doing solely this).

Structure prediction

Proteins are transcribed genes. Although proteins are fully described by their amino acid sequence, the binding process depends largely on the 3D physical structure that a protein adopts. Numerous groups are active in predicting these 3D structures from the primary amino acid sequence description.

Function prediction from sequence data

Since genomes are becoming available more frequently and together with them descriptions of genes, the aim of function prediction is to discover the function of those genes within the genome. The usual approach to identifying the function of an unknown gene is — roughly — to compare its sequence information (again the C, T, A, G's) to the description of known genes. If such 'homologous' genes are found, then the function of the unknown gene is derived from the function of the known homolog. Naturally, the sequence information is not exact, so search algorithms are developed that can cope with partial mismatches. Another challenging factor is that such sequence information is stored in numerous databases throughout the world. Besides searching such distributed databases, other aspects such as consistency, data validity etc. play a role.

Function prediction from gene activity levels

A new area in bioinformatics is the discovery of functions of genes and interrelationships between genes, not from the sequence information, but from the *measured* activity levels of the genes (gene expression levels, mRNA). These measurements are taken under a variety of conditions (e.g. heat shock, oxygen restriction etc.). The expression levels of the cell can be measured statically (before and after) or even dynamically (time series measurements). Another possibility is to measure the expression levels of a cell that is mutated, i.e. where one of the genes is removed or copied, when subjected to these conditions. The objectives of these bioinformaticians is to infer from these input-output relationships of the cell, *new* knowledge about this system in terms of gene functions or relations, and how genes influence each other in particular pathways. Such information can be derived from the measurements by looking at co-regulated behavior or more complex networks of regulatory behavior.

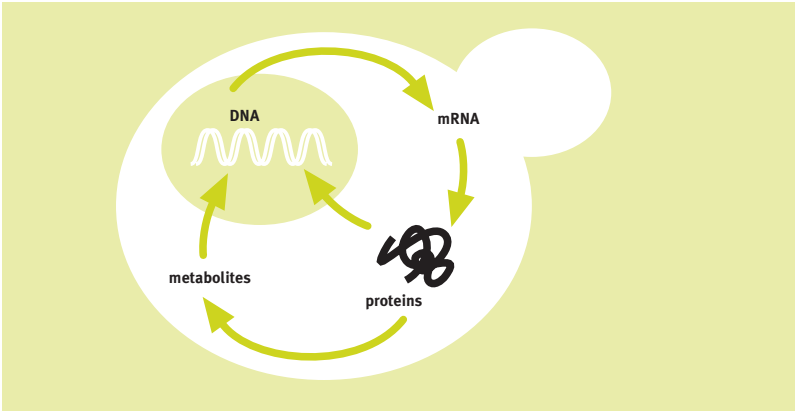


Figure 2
The cell as a dynamic system.

Data mining within bioinformatics

Characteristics of the available data

Types of data

Basically two types of data are available: sequence data and expression data. Both types of data are available in the 'DNA' (genome) and 'amino acid' (proteome) domain. This implies that one could have 1. genomic sequence information (ordering of G, C, A and T on the genome), 2. protein sequence information, i.e. ordering of amino acids in a protein, 3. gene expression data, i.e measurements of gene activity as well as 4. protein expression data, i.e. measurements of protein concentration and activity levels. In addition, measurements of metabolite concentration levels are also available. In some cases, protein sequence data is augmented with structural data, which reveals the 3D structure assumed by a particular amino acid sequence (protein).

Quality and quantity

Genomic sequence data is probably the most abundant type of data. This stems from the fact that the human genome project resulted in a tremendous technological revolution in sequencing and database technology. Consequently, it has become fast and relatively cheap to sequence genomes. In contrast, the technology to determine protein sequences (Liquid Chromotography-Mass Spectrometry-Mass Spectrometry, LC-MS-MS) has a smaller throughput. This will probably change in the near future, as the interest gradually shifts from genes to gene products. Gathering 3D protein structure information is extremely time-consuming. Consequently, this information is only available for a small number of proteins.

The human genome project also accelerated the development of technologies (micro arrays) for the measurement of gene expression levels. As a result, there has been an explosion in the amount of available gene expression data. It should be noted that gene discovery, i.e. identification of the coding regions in the genome is a precursor to the development of micro arrays that measure the activity levels of genes. This is one reason why expression data is less abundant than the primary sequence data. An additional factor, which still limits the availability of expression data, is cost. Since a genome usually contains thousands of genes, these data sets only become useful when the gene expression is also measured under a wide range of conditions such as different mutations, exposure to different drugs or time points. The cost associated with each measurement is still prohibitive (especially in academic settings) and thus compromises the usability of the data. This will probably change in the near future, as the technology is more widely used. In spite of these drawbacks, mining of existing gene expression data can already produce valuable insights into genetic regulatory mechanisms.

Protein expression data is also far less commonly available than gene expression data. Protein chips are being developed, and will probably become widely used within the next five years.

Ownership and accessibility

Most of the genomic and proteomic sequence data is publicly available. However, information derived from this primary data, i.e. gene and protein function information, is often proprietary. This is caused by the fact that pharmaceutical companies perform a major part of this research, and base their drug development programs on information about gene and protein function (specifically with respect to diseases). The same holds for expression data, where compendiums of expression data (genome wide expression measurements for a large number of drugs and conditions) remain proprietary.

Possibility of experimenting (to gather data)

Gathering biological data through experiments obviously requires specialized equipment and personnel. It is not an activity undertaken by bioinformaticians on a regular basis. Such data is usually acquired from public sources, or through collaborations. The latter is the most preferable setting, since close collaboration with biologists also provides access to their expertise, which is invaluable for practicing meaningful bioinformatics.

What questions are we trying to answer?

The holy grail of bioinformatics basically boils down to figuring out how living organisms (or subsystems within an organism) work, in order to correct them in cases of abnormality (fighting disease), or to engineer them in such a way that they produce the desired behavior (e.g. 'microbial factories' which produce medicine). More specifically, this involves the determination of the function of the different players in the metabolism (genes, proteins and metabolites) and how they interact. The approaches towards answering this question are dictated by the data used as starting point. Here we outline typical questions for different types of data. These questions are ordered hierarchically: from basic to more complex questions and are not intended to be exhaustive, but to give an impression of the types of questions being addressed.

Genomic sequence data

- Where are the genes in the genome?
- What are the functions of these genes?
  - Which promoter sites are associated with this gene?
  - Is this gene homologous to other genes of known function?

What protein products does this gene produce, and can the gene's function be inferred from these products?

- Comparative genomics: How does a particular genome compare to other known genomes, for example, with relation to the preservation of some vital cellular functions, pathways etc.?

Protein sequence data

Which protein family does this protein belong to and does it shed light on its potential function? (Finding homologous proteins)

- What is the 3D structure of this protein given its sequence and what can we derive from it about its function?

3D protein structure information

- What are potential binding sites on this protein?
- What is the function of a protein, given its sequence and 3D structure? (What other proteins/genes/metabolites does it bind to, i.e. interact with?)

Gene expression data

Which genes are co-regulated, i.e. have similar expression profiles?
How are genes interacting, i.e. which genes activate or inhibit a particular gene or set of genes?

Data mining steps in bioinformatics

As in many other science areas the first step is hypothesis generation, e.g. to identify different cancer types, a starting point could be a current taxonomy of cancer based on current know-how. This is followed by an experimentation step, i.e. the collection of data to improve the current hypothesis. In the case of the example, this amounts to collecting gene expression data for a variety of cancer types and conditions. The next step will be model building (knowledge discovery) based on 1. the generated data, 2. existing data, 3. existing knowledge and 4. the current hypothesis. In the example, this amounts to refining the taxonomy based on clustering of expression profiles into distinct classes or types. To verify the new hypothesis, new experiments are conducted to check the class or type predictions, and, if necessary, enter a new cycle by performing new (more specific or refined experiments), updating the model and verifying. This process is continued until a satisfactory model has been obtained.

State of the art (for data mining applications in the area)

Gene prediction based on sequence data

Approaches to gene finding are based on:

- the structure of typical genes, i.e. specific arrangements of gene subunits such as promotor sites, introns as well as exons;
- the typical characteristics (nucleotide make-up) of these subunits.

First, global statistics are employed to determine likely locations of genes. This excludes the inter-genic regions, i.e. regions where genes are unlikely to be located. Given these likely locations, each location can be analyzed in more detail with, for example Hidden Markov Models (HMM, see Section 6.2.10). HMMs can model the make-up of a genes as match states in the model. For gene finding, a hidden Markov Model could have four major match states that represent the promoter, the start codon, splice sites and the stop codon. Other match states can be employed to represent more subtle statistical patterns (such as transcription factor binding sites, or enhancers) that occur before, between or after these patterns. Most methods concentrate on finding a particular subunit in a gene. See [Bajic, 2000] for an overview of different approaches.

Protein function prediction

The most effective way to obtain information about protein function is to obtain an accurate as possible description of the 3D structure of the protein, since this structure dictates the function of the protein: what it binds to and where it binds to it. There are several steps that are typically followed to determine protein function.

The most obvious first stage in the analysis is to perform comparisons with sequence databases to find proteins with similar sequences, i.e. homologs. There are numerous web servers for doing searches, where one can post or paste a sequence into the server and receive the results interactively. There are many methods for sequence searching. The most well known is the BLAST[4] suite of programs. One of the most important advances has been the development of both gapped BLAST and PSI-BLAST (position specific integrated BLAST). These adaptations have increased the sensitivity of BLAST.

After a database search, a multiple sequence alignment containing the given protein sequence and all the found homologs is performed. Probably the most important recent advance has been in the Hidden Markov Models (HMM) approaches for sequence alignment.

If no homolog with known 3D structure can be found, a logical next step is to attempt to predict or approximate the 3D structure. This is a complicated process, and reliable 3D structure prediction is fairly complicated. It involves (a possible combination of) several steps, including secondary structure prediction [Wako, 1994; Mehta, 1995; King, 1996] fold family analysis [Lemer, 1996], tertiary structure prediction, alignment of secondary structures and comparative modeling.

Function prediction from expression data

These approaches are also referred to as 'reverse engineering' of regulatory genetic, protein and metabolic networks based on expression data. Basically two major approaches exist.

Static approaches

The most important static approach is clustering. In this approach, similar expression profiles are grouped together under the assumption that genes with similar expression profiles also have similar functionality. See e.g. [Eisen, 1998; Michaels, 1998]

In classification approaches, function is determined by mapping expression profiles to a set of predefined functional classes. To achieve this, a classifier is trained on a set of labeled data. Examples of this approach can be found in [Golub, 1999].

4 http://www.ncbi.nlm.nih.gov/blast/

Dynamic approaches

In these approaches, relationships between genes (or proteins and metabolites) are derived by fitting a parameterized model to time series (i.e. dynamic data). The parameters of the model provide information about possible interactions between the entities included in the model. For example, in linear genetic network models, a large positive weight is interpreted as an activation function, while a large negative weight implies an inhibitory interaction. Several different types of genetic network models have recently been proposed. These include Boolean networks [Liang, 1998], Bayesian networks [Friedman, 1999a; Friedman, 1999b], (Quasi)-Linear networks [Wessels, 2000; Someren, 2000; D'Haeseleer, 1999], Neural networks [Marnellos, 2000] and Complex Models [Chen, 1999].

Future

In the future technology to measure expression levels will become widely used, i.e. the amount of expression data will grow significantly. Consequently technologies to unravel the combinatorial regulatory interactions will become more prominent, as well as database technology to represent these relationships. Interactive tools allowing scientists to ask 'what-if' questions will also become more widely used. Protein technology will expand and develop as genomic technologies become more and more established. The link between genes, proteins and metabolites will also become more prominent, i.e. bioinformatics will become a science that provides a truly integrative perspective on biological systems.

References

- Bajic, V.B. (2000). Comparing the Success of Different Prediction Software in Sequence Analysis: a Review. Briefings in Bioinformatics **1** (3):214-228
- Chen, T., H.L. He, G.M. Church. (1999). Modeling Gene Expression with Differential Equations. Pacific Symposium on Biocomputing '99. Vol. **4**:29–40
- D'Haeseleer, P., X. Wen, S. Fuhrman, R. Somogyi. (1999). Linear Modeling of mRNA Expression Levels During CNS Development and Injury. Pacific Symposium on Biocomputing '99. Vol. **4**:42–52
- Eisen, M.B., P.T. Spellman, P.O. Brown, D. Botstein. (1998). Cluster Analysis and Display of Genome-Wide Expression Patterns. Proceedings National Academy of Science, USA 95. pp14863-14868
- Friedman, N., M. Goldszmidt, A. Wyner. (1999a, 1998). Data Analysis with Baysian Networks: A Bootstrap Approach. Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI)
- Friedman, N., M. Linial, I. Nachman, D. Pe'er. (1999b). Using Bayesian Networks to Analyze Expression Data. Submitted for publication

- Golub, T.R., D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science **286**:531-537
- King, R.D., M.J.E. Sternberg. (1996). Identification and Application of the Concepts Important for Accurate and Reliable Protein Secondary Structure Prediction. Protein Science **5**:2298-2310
- Lemer, C., M.J. Rooman, S.J. Wodak. (1996). Protein Structure Prediction by Threading Methods: Evaluation Of Current Techniques, PROTEINS: Structure, Function and Genetics **23**:337-355 (Assessment of Techniques)
- Liang, S., S. Fuhrman, R. Somogyi. (1998). REVEAL, A General Reverse Engineering Algorithm for Inference of Genetic Network Architectures. Pacific Symposium on Biocomputing '98, Vol. **3**:18–29
- Marnellos, G., G.A. Deblandre, E. Mjolsness, C. Kintner. (2000). Delta-Notch Lateral Inhibitory Patterning in the Emergence of Ciliated Cells in Xenopus: Experimental Observations and a Gene Network Model. Pacific Symposium on Biocomputing 2000, Vol. **5**
- Mehta, P., J. Heringa, P. Argos. (1995). A Simple and Fast Approach to Prediction of Protein Secondary Structure from Multiple Aligned Sequences with Accuracy above 70. Protein Science **4**:2517-2525
- Meyers, G. (2000). http://www.the scientist.com/yr2000/may/prof_000501.html
- Michaels, G.S., D.B. Carr, M. Askenazi, S. Fuhrman, X. Wen, R. Somogyi. (1998). Cluster Analysis and Data Visualization of Large-Scale Gene Expression Data. Pacific Symposium on Biocomputing '98. Vol. **3**:42-53
- Pandey, A., M. Mann. (2000). Proteomics to Study Genes and Genomes. Nature **405**:837-846
- Someren, E.P. van, L.F.A. Wessels, M.J.T. Reinders. (2000). Linear Modeling of Genetic Networks from Experimental Data. Accepted for presentation at ISMB 2000
- Wako, H., T.L. Blundell. (1994). Use of Amino-Acid Environment-Dependent Substitution Tables and Conformational Propensities in Structure Prediction from Aligned Sequences of Homologous Proteins. Part 2. Secondary Structures. Journal of Molecular Biology **238**:693-708
- Wessels, L.F.A., E.P. van Someren, M.J.T. Reinders. (2000). Analyzing Gene Expression Data. Proceedings of the XX International Congress of the International Society for Analytical Cytology (ISAC 2000). Montpellier, France

2.3.4 Data mining for genomics and drug discovery

Luc Dehaspe[1]

Introduction

In general terms, genomics concerns the effort to identify the genome, the full set of instructions for the construction, (mal-) functioning, and death of an organism.

Some landmark discoveries in genomics include: the theory of evolution by natural selection (Darwin and Wallace, 1858), the presence of heritable traits — we would now call genes — in peas (Mendel, 1866), the helical structure of DNA (Franklin, 1951), the three-dimensional double helix structure of DNA (Watson, 1953), and a method for sequencing DNA (Sanger, 1977). More recently, genomics research has gained momentum through the competition between the public 'Human Genome Project' and Celera Genomics leading to the simultaneous parallel publication of two 'first drafts' of the human genome in Nature (Vol. 409, 15 February 2001) and Science (Vol. 291, 16 February 2001).

As the full picture of an increasing number of genomes — including our own — is completed, attention shifts towards the interpretation of that picture. The impact of such knowledge derived from genomic data can hardly be overestimated. It could provide insight into the mechanisms of disease and cure. To the pharmaceutical industry genomics holds the promise of triggering new drugs that increase the quality of life of numerous new patient populations. In this chapter we will look at genomics from that pharmaceutical perspective and discuss contributions of data mining to the drug discovery area, where genomics data join the flood of chemical, clinical, and other biological data.

One might reasonably expect data mining would have been invented for genomics and drug discovery, if the data flood had not been a more general phenomenon. In fact, a significant part of the data analysis technology in this application domain did develop independently. This will become clear in the following paragraphs, as we detail some of the specificities of genomics, characterize the available data, and sketch the state of the art for data mining applications.

Specifics of the drug discovery process

Target discovery

The role of genomics in the drug discovery process is to identify targets for drugs. Drugable targets include substances such as enzymes and receptors, and processes such as gene expression and signal transduction. As far as a target plays a role in causing a disease, modulating the target chemically to

1 Dr L. Dehaspe, Luc.Dehaspe@cs.kuleuven.ac.be, PharmaDM, B-3001 Heverlee, Belgium, http://www.PharmaDM.com

restore its function to near normal might stop the related disease process. Only a decade ago the number of targets was limited to about 500. In the following years this number is expected to increase to 10,000.

The first step in target discovery is to sequence the genome[2], the next is to *find the genes*. Particular subsequences of DNA encode for proteins, the laborers that are responsible for the interplay of biochemical processes called life. In case the protein has been observed, the gene that produced it can be localized on the DNA string. Using these protein-gene couples as examples, scientists have induced characteristics of protein-coding stretches of DNA. On the basis of these rules hypothetical — unobserved — proteins have been proposed.

Even with full information on all proteins — the so-called proteome — of an organism, we are left with the key question: What do these proteins actually do? This question is addressed in functional genomics.

One strategy to find out about the function of a gene is to determine its three-dimensional structure. What is true for any tool also holds for proteins: its shape might provide a hint as to what it is used for. For a limited number of proteins the 3D structure has been determined experimentally via crystallography. Other laboratory experiments carried out to discover the function of proteins involve modifications to the genetic material of organism that block the production of a protein. Once the protein is gone, one might be able to observe which processes are affected by its absence. Both structure determination and knock out experiments tend to consume enormous amounts of resources and time.

The techniques above start from a particular protein for which the function is determined. Alternatively, one might start from a particular function and look

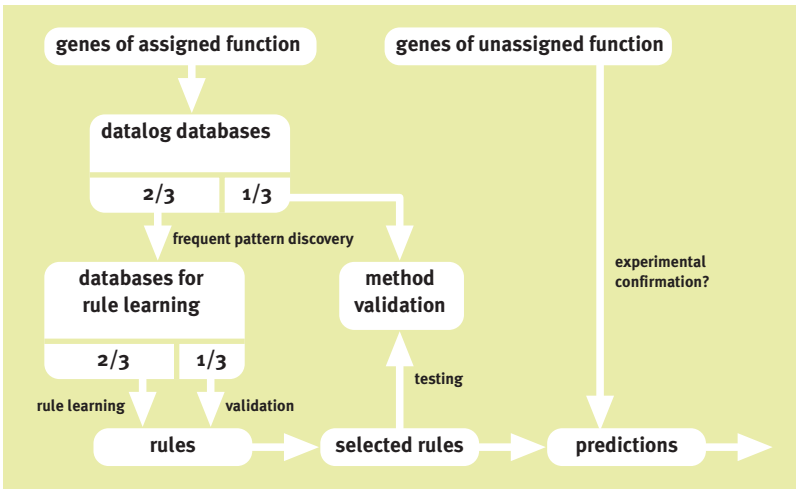


Figure 1
Flow chart of the experimental methodology followed in the functional genomics application.

2 The technologies used in this data generation stage are beyond the scope of this volume. Details can be found in the Nature and Science special 'Human Genome' issues cited above.

for the proteins involved. This is the basic idea underlying micro array experiments, in which all the genes of an organism are represented on a microscope slide. Microchip technology allows researchers to estimate the amount of proteins produced in particular cell types and biological processes. For instance, one might compare protein levels in healthy tissue and diseased tissue. Proteins where these levels differ significantly can be held responsible for keeping the cell healthy or for sustaining the diseased state.

Drug discovery

Once a valuable drug target has been identified it still takes a pharmaceutical company about 6 years and 200 million Euro to come with a candidate drug.

A drug is typically a small molecule that binds to the target protein much as a key fits a keyhole. If you know the three dimensional structure of the keyhole well enough, it makes a lot of sense to build the perfect key from scratch using that information. In the other case, one has to fall back on a less rational approach and start screening large libraries of chemical compounds for candidates that loosely fit. In such high throughput screening experiments a database is generated with a lot of implicit information on relationships between chemical structure of compounds and their biological activities.

The next step is to enter a selection of acceptable candidates — called leads — into a polishing phase. The purpose of this 'lead optimization' process is to preserve the desirable binding properties of selected molecules while suppressing the unwanted properties. A valuable drug not only binds to the target in vitro, but is also capable of reaching that target in vivo, has the desired biological effect at low doses, has no harmful side-effects, is safely excreted afterwards, can be administered orally, and can be patented and produced, shipped, stored, and sold in great numbers.

Clinical trials

A candidate drug that comes out of the lead optimization stage still has to go through at least 7 years of clinical trials during which it is tested on humans. The idea is to find out about the appropriate circumstances, if any, in which this drug might be prescribed.

Conclusion

Developing a typical small drug takes roughly 15 years and costs 300-600 million Euro. It should therefore be no surprise that technologies such as data mining that might speed up the drug discovery process are perceived as business-critical by the pharmaceutical and biotechnology industries.

Characteristics of the available data

Quantity

It is hard to perceive an upper boundary on the quantity of data that can be generated in target discovery, high throughput screening of drugs, and clinical trials.

On the target discovery side, a classical example is the yearly doubling size of GenBank, the genetic sequence database of the U.S. National Institute of Health. The GenBank repository contained more than 12 million sequences as of June 2001. In the drug discovery stage, ultra high throughput screening technology allows companies to screen over a hundred thousand compounds a day. Finally, clinical trials are continued as long as a drug is on the market. Consequently, data on the reactions of patients to a particular drug continue to accumulate.

Quality

Drug discovery data have often been generated in a laboratory context. One aspect of data quality concerns the experimental setup: to what extent do we know what we are measuring? This question becomes more difficult to answer as we move from in vitro tests involving one protein and one drug to cyto tests involving a cells or tissues to in vivo tests where we measure some biological response in living organisms (e.g. rodents or humans). As we ascend the in vitro to in vivo scale, experiments become more relevant, but less controlled and reproducible.
Obviously, the quality of the measuring instruments has a major impact on data quality. Recently, a micro array vendor had to withdraw a set of microchips on which the genes were organized in the wrong orientation. It is telling that it took their clients some time to realize their measurements were totally useless.

Data quality is even more problematic as we rely on public domain data where the exact experimental setup is often not specified and the scientific authority of submitters is not guaranteed. For instance, GenBank is known to contain many useless entries, and even plain junk.

Type

In the pool of biological, chemical, and clinical data described before you will encounter the full spectrum of data types. Obviously, numerical data abound: e.g. gene expression levels, 3D coordinates of atoms, binding affinities, toxic and active doses. Symbolic data include atom bond structures of molecules, taxonomies of species, functional classes of proteins, and descriptions of patients. The database also contains unformatted data: free text documents, images and movies.

Apart from great variety in individual data points, drug discovery data also show complex relationships between those data points. An obvious example concerns the relations within a molecule between atoms and bonds. Similarly, a single protein, drug, or patient will be involved in a set of experiments. It should then be clear that the relational database used to store drug discovery data contains many tables connected by numerous many-to-one relationships. Mining this complex database efficiently will require tools that can cope with multiple relations.

Role of data mining

Target discovery data

A useful overview (with references and web links) of the data mining tasks involved in target discovery can be found in [Mount, 2001]. We briefly introduce a selection.

Given examples of subsequences of DNA known to code for proteins, build a classifier that is able to recognize such sequences.
Subtasks of this task include: finding promotor sequences that precede a protein coding sequence; in the protein coding DNA string, discriminate between the exons and the introns, i.e. the parts that will be translated to part of a protein and those that will be deleted before translation.

Given examples of proteins and their functional class, build a classifier that can assign function to new proteins.
This task is conventionally addressed using nearest neighbor methods. From a database such as GenBank those sequences are retrieved that show significant sequence similarity. The assumption is that since the structure of these proteins has been preserved by evolution, they may also share functional characteristics. Special purpose sequence similarity searching algorithms such as FASTA [Pearson, 1988] and PSI-BLAST [Altschul, 1997] are amongst the most popular software tools in biology today.
Alternatively, rule based data mining approaches can be used to induce rules which map from sequence to functional class, as done in [King, 2000; King, 2001]. This publication illustrates the advantages of relational data mining approaches in applications where data are inherently relational and where domain specific background knowledge is available. The flow chart of the experimental methodology is shown in Figure 1. The relational data mining algorithm Warmr is used to discover frequent patterns in the datalog databases built from two bacterial genomes. Only 2/3 of the database was used in this step; the remaining third was set aside as the final test set. Using the frequent patterns as binary features, a new — single table format — database was constructed. In

this database data mining algorithm C5 discovered rules that predict function from the descriptional attributes using 4/9 of the data (2/3 of the 2/3). Good rules were selected on the validation data — the remaining 2/9 of the dataset. The unbiased accuracy of these rules was estimated on the 1/3 test set. This methodology resulted in good rules that predict function from sequence. The test accuracy of these rules was far higher than possible by chance. Of the genes of no assigned function 65% were predicted to have a function. For instance, the rule

If the percentage composition of lysine in the gene is › 6,6%
Then its functional class is 'Macromolecule metabolism'[3]

turned out to have an accuracy of 11/13 (85%) on the test set, where the probability of this result occurring by chance is approximately 10^{-5}. This rule predicted class 'Macromolecule metabolism' for 15 genes of unknown function. It illustrates the fact that many of the predictive rules were found to be more general than possible using sequence homology: they correctly predict the functions of proteins (1) from more than one homology grouping and (2) not homologous to any in the training data.

Given examples of proteins and their secondary structure, build a classifier that predicts this structure.
This problem has received a lot of attention in bioinformatics. The best current solutions combine statistics and data mining and sequence similarity technology.

Given results of micro array experiment, cluster those genes that respond the same way to an environmental signal.
This is probably the drug discovery task that receives most attention in the data mining community. Hierarchical clustering techniques especially have been widely applied [Eisen, 1998].

Drug discovery

Given examples of molecules and their biological activity, relate chemical structure to activity and predict activity of new compounds.
Traditionally this problem has been addressed with linear regression, more sophisticated statistical methods, neural networks, and feature-based data mining techniques. All these approaches attempt to capture the key features of molecular structure in a feature vector representation. Based on a more expressive language, relation data mining approaches allow discovery of relationships that are inaccessible to conventional methods and output these discoveries in a language easily understood by chemists [King, 1996].

3 This rule is consistent with protein chemistry: lysine is positively charged which is desirable for interaction with negatively charged RNA.

Clinical data

Given examples of drugs and their effect on patients with a particular profile, build a model that predicts the effect of a drug on a patient.

A wide range of data mining algorithms can be applied to this task. A related problem concerns prediction of therapy-response. For instance, HIV-infected patients tend to develop resistance to particular drugs and optimal, individualized therapies have to be proposed.

Vision of the future

We foresee integration will become a key theme at four distinct levels.

Integration of chemical, biological, and clinical data

Currently the three stages in drug discovery are typically studied separately, often in separate departments in pharmaceutical companies. There is a growing awareness that each of these stages could benefit from the inclusion of data and knowledge generated at the other stages. One example of where this integration could be realized is pharmacogenomics, the study of how to tailor drugs to individual genetic profiles of patients.

A limiting factor will be the flexibility of data mining technology to handle data that is not stored in a single table (spreadsheet like) format, but rather in rich relational (or object-oriented) databases. Relational data mining technology anticipates this evolution.

Integration of numerical and or statistical and symbolic data mining

As in many data mining applications, success criteria might include both comprehensibility and predictive accuracy of the model. The output of numerical and statistical techniques is often hard to understand. Although symbolic methods have an advantage on that front, optimal predictive performance typically requires a mixture of many methods. A complete solution should therefore be flexible and allow the user to trade understandability for accuracy.

Integration of data mining experts and drug discovery experts

Life sciences applications have always been popular in the data mining community. All too often however these applications only serve to demonstrate the benefits of particular data mining technologies. A simplistic outsider's view on the application domain cannot lead to significant contributions to that domain, especially not when that domain has been studied extensively by generations of leading scientists.

Conversely, the drug discovery community tends to have a limited view of data mining solutions, typically based on the methods used in seminal publications. One can expect a tendency to realize the full potential of data mining in drug discovery in multidisciplinary teams.

Integration of data generation equipment and data analysis software

At the moment data mining takes up a passive role in its relationship with laboratory equipment. One can envisage a more active participation of data mining in the drug discovery process where it would assist in monitoring the quality of data and initiate additional experiments required to choose between competing hypotheses. The latter idea is already being explored in a 'Robot Scientist' project in the United Kingdom.

References

- Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman. (1997). Gapped BLAST and PSI-BLAST: a New Generation of Protein Database Search Programs. Nucleic Acids Research **5:**3390-3402
- Eisen, M.B., P.T. Spellman, P.O. Brown, D. Botstein. (1998). Cluster Analysis and Display of Genome-Wide Expression Patterns. Proceedings National Academy of Science **95**:14863-14868
- King, R.D., S.H. Muggleton, A. Srinivasan, M.J.E. Sternberg. (1996). Structure-Activity Relationships Derived by Machine Learning: the Use of Atoms and their Bond Connectivities to Predict Mutagenicity by Inductive Logic Programming. Proceedings National Academy of Science **93**:438-442
- King, R.D., A. Karwath, A. Clare, L. Dehaspe. (2000). Accurate Prediction of Protein Functional Class in the *M. tuberculosis* and *E. coli* Genomes Using Data Mining. Yeast (Comparative and Functional Genomics) **17**:283-293
- King, R.D., A. Karwath, A. Clare, L. Dehaspe. (2001). The Utility of Different Representations of Protein Sequence for Predicting Functional Class. Bioinformatics **17**:445-454
- Mount, D.W. (2001). Bioinformatics. Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press, New York
- Pearson, W.R., D.J. Lipman. (1988). Improved Tools for Biological Sequence Comparison. Proceedings National Academy of Science **85**: 2444-2448

2.3.5 Mining museum riches

An introduction to digital methods in the discovery of changes in the diversity and distribution of Life on Earth.

Jan Krikken[1], *William H. Piel*[2]

Introduction

Data mining (or knowledge discovery in databases, KDD) has been defined as the non-trivial extraction of implicit, previously unknown, and potentially useful information from data (this volume; [Frawley, 1992]. When the question was posed to us: "are you, in your museums, doing any data mining, and if so, what can be expected of this in the coming decades?" our first response was: "are we as research curators in natural history institutions, doing anything else?" Thus, before discussing data mining itself, we should introduce the subject that concerns us here: the study of biodiversity and the generation of data in the context of natural history museums.

Scope of this paper

Although this paper is intended primarily to introduce non-specialist data miners into a number of aspects related to biodiversity collection facilities, institutions in certain other disciplines (particularly in archaeology and ethnology) may well prove to be fundamentally similar. Collections of cultural diversity may differ in the spatiotemporal scale of the evolution of their ingredients – their creators usually being short-lived humans, but otherwise much of what is said in this paper applies to them as well. The volume of data in cultural history collections and the inherent information management aspects have, to our knowledge, not been assessed globally as will be done further below.

At any rate, in all collection diversity studies, data mining is conditional on collection mining, which is conditional on real world sampling. Consequently the workflow stages of accumulating and studying museum collections, data entry activities, and their inherent pitfalls, are briefly reviewed hereafter. [Frawley, 1992] refers to specific database-related problems, which are all too familiar to users of biodiversity databases, like missing values, missing fields, errors, and noise; these are not re-discussed here.

After the introduction to workflow stages follows a selection of examples from a broad range of biodiversity pattern studies, ranging from straightforward statistical database analysis to interactive pattern classification and identification. Work on living organisms is complementary to work on preserved objects, and one particular domain worth mentioning is the development of biomolecular databases (GenBank and the like). These are expanding extremely rapidly

1 Drs J. Krikken, krikken@nnm.nl, National Museum of Natural History, Museum Naturalis, Leiden, The Netherlands.

2 W.H. Piel, piel@rulsfb.leidenuniv.nl, National Museum of Natural History, Museum Naturalis, Leiden, The Netherlands.

[Baxevanis, 1998], and have also had input from the world of 'ancient DNA' in the last decade, such as sequencing tissues from collections of extinct organisms [Herrmann, 1994; Hofreiter, 2001].

Museum collections harbour information about past and present biodiversity change

Collections from nature have always, and particularly since Linnaeus developed a nomenclatural standard (e.g. [Linnaeus, 1753, 1758], been the subject of intense scientific scrutiny. These collections, if not materially or intellectually valuable for their own sake, are always carriers of interesting, non-trivial information. Linnaeus already knew this, the economic use of plant diversity being high on his agenda [Koerner, 1999]). With Linnaeus the encyclopaedic representation of nature's diversity began in earnest; and in the present information age it continues at a pace as never before (see [Levin, 2001] for an extensive overview, [Hancock, 2000] for complementary Earth science aspects). The mountain of data meanwhile extracted from natural objects kept in collections (naturalia, for short) is enormous, and much more can be extracted in view of the endeavor to discover and understand patterns of evolutionary and ecological change. Data extracted from collections mainly deal with the morphology of the naturalia and their distribution in space and time — in short: their evolutionary history. Institutional collections have, in terms of numbers of specimens, only rarely been assembled to answer specific research questions. Usually, they have not even been collected for spatiotemporal mapping, as the majority of specimens and relevant data stem from non-professional collectors and recorders. For certain groups of organisms (for instance, insects, which make up two-thirds of the world's known species) the share of specimens accumulated by amateur collectors may be over 70%, particularly in Europe. Collecting purely in the interest of building up a collection for its own sake, without sophisticated research objectives, means that any future data pattern analysis should fall under the above definition of data mining (or KDD). In recent years biodiversity has been recognized as a globally valuable resource (National Research Council, NRC, 1999), a recognition formalized through international conventions like the Convention on Biological Diversity, CBD, and the Convention on International Trade in Endangered Species of Fauna and Flora, CITES. As a result, the inherent voluminous information dimension has at last also attracted the attention of politicians and administrators, the Global Biodiversity Information Facility of the OECD being a case in point ([GBIF, 2001], see also International E-networking).

Central role of classification and naming in discovering biodiversity patterns

The efforts of Linnaeus and numerous taxonomists after him are of crucial importance in studying the diversity of Life on Earth. Taxonomists have provided the hierarchic system of name tags connecting any biodiversity data to particular groups and, ultimately, to species of organisms, alive and extinct. In spite of the recent debate on the usefulness of the current, essentially Linnaean system of bio-nomenclature (e.g. [Minelli, 1999; Pennisi, 2001; Kress, 2001]), the identification of organisms depends on this name-tagging system totally, and will undoubtedly do so for some time to come. The mining of species-related data is crucially dependent on a stable nomenclatural, terminological and conceptual system. The system should be warranting that we all know, and agree upon, what qualitative and quantitative content exactly we are looking for. Certain diversity studies (such as those on the ecosystem level), though they may not be directly dependent on a complete nomenclaturally correct identification of all the organisms involved, still assume a systematic framework of operationally recognizable species, or groups of species. Recently, work has been invested in setting up a species-level web portal to access a multitude of web-based biodiversity content providers [Species 2000, 2000].

Retrospective and immediate biodiversity data entry

Vast amounts of biodiversity data are stored almost exclusively in numerous shelf kilometers of monographs, other types of synoptic work, and a whole range of other hardcopy documents (scientific papers, archival documents, in-house collection registers, card indexes, label data, etc.). These documents and their data are now gradually finding their way into different types of digital databases — with data quality as a notable background problem. Examples of such databases are global group-delimited systems, like FishBase and numerous others, which may be consulted through an umbrella search engine (Species 2000), and various regional or national systems under development (e.g. Fauna Europaea, International Plant Names Index, etc.). Nowadays both the primary event data coming with newly collected species and the secondary data extracted from the systematic study of extant collections are finding their way straight into digital databases, be they simple tables in office applications (like Microsoft Access) or more sophisticated systems (applications in SQL-server based systems).

What to archive for future generations: the futurological dilemma

The information enshrined in the physical collection objects (in the museum world the general term realia is frequently applied) is manifold and voluminous (see below). The museum and herbarium taxonomist is primarily interested in the diversity of the attributes of the objects for use in identification, classifica-

tion, and biogeography. Biogeographic mapping, as in species distribution maps, plays a central role, but many other information categories may, sooner or later, prove to be of significance. This could be in another context, as in the study of symbiotic associations, or on different environmental scales, such as with species-soil relationships. Relevant data may, even unwittingly, have been recorded or physically collected in the field, although they might only be of no direct significance to the collector or recorder. Fundamentally speaking, the information that may be required in a future analysis is never certain — this is the futurological dilemma in the acquisition and management of collections of realia and associated data of any kind, and, for that matter, in archival operations in general. This fundamental uncertainty is precisely the rationale of subsequent data mining. It is also the reason why archival data storage systems are not merely continually updated databases, but take the form of data warehouses, i.e. database systems in which the historical dimensions of both the content and procedural data are fully preserved.

Data categories: image patterns and the issue of the so-called virtual museum

The data need not necessarily be limited to qualitative and quantitative data in the stricter sense. They may take the form of images, which may, given the right tools, also be searched and analyzed. As the study of the diversity of Life on Earth is a highly visual business, the development of imaging software for recording and subsequently mining diversity is likely to become pivotal in the years ahead. The identification of images of entire organisms, of body parts, of particular attributes, or whatever, is not new. Image, or more generally, pattern recognition is already integrated in various software applications, and, although biodiversity applications are still few in number; they will come. It should be noted here that virtual historical collections (say, data warehouses with an emphasis on image storage) can, in view of the futurological dilemma, never replace collections of realia, as is sometimes rumored in circles of policy makers and museum directors, and prompted by economic motives.

Volume and nature of biodiversity data

Some reflections on data volume, complexity, and reliability in biodiversity research are in order [Groombridge, 1992; Minelli, 1993]. Supposing that there are 200 larger, well-curated collection facilities worldwide, each with 5-10 million individual naturalia (the natural history collections of the Smithsonian are said to include 120 million, apart from Paris and London most national collections in Europe contain 10-20 million). There are thus around 1-2 billion objects kept in natural history collections world-wide, belonging to roughly 1,5 million known species. A small country like The Netherlands has records for somewhere around 36,000 species [Nieukerken, 1995], and the actual amount may

well rise to 50,000 species. The vast majority of organisms ('the other 99%', also called cryptobiota) are known almost entirely from collections of preserved material. Conservatively estimated, there are in the order of ten million species of living organisms world-wide, each with hundreds of defined and still definable attributes. Fossil species constitute an additional quantity, possibly not very large (many groups of organisms fossilize poorly), but, at any rate, still poorly known. Even with these seemingly large figures, the geographic distribution of at most the vertebrates, butterflies and higher plants of Europe, North America, and possibly Australia and a few parts of the world elsewhere, can be reliably mapped. Considering the volume of data, it is easy to see that the ultimate product will reach quantities that can only sensibly be managed and mined by defining one's universe of discourse clearly, and by concentrating on the development of generic tools that fit a given discourse. The global approach is discussed further below. In dealing with collections of realia, data mining has a special connotation. Delving into the collections and extracting data of evolutionary and historical significance from their components is indeed the primary business of the research curator in natural history facilities, or, for that matter, any type of archival collection facility. Data mining is simply part of the curator's ethos.

Work processes and the information cycle

An outline of the work flow in the mainstream discipline of biodiversity collection studies, taxonomy [Winston, 1999] is a useful practical introduction, can show the various data categories generated in the successive stages, the pattern analysis of the databases, and the presentation of interesting patterns discovered. Most data categories generated during the taxonomic work flow are nowadays directly or retrospectively entered into digital databases. Distinguishing, naming, describing, and classifying taxa (singular: taxon), i.e. species and other units of hierarchical classification, is the main task of present day taxonomy, as it was in Linnaeus' time. In a broader scientific context these alpha stages are usually followed by attempts to reconstruct the evolution of the taxa and the mechanisms involved. This evolution is usually represented by an evolutionary tree, a partial 'Tree of Life' (a phylogeny), the branches of the tree representing the diversification processes forming the taxa. The knowledge gained from advances in taxonomy and phylogeny work their way back into the initial identification and curation tasks performed by the museum staff, and hence the information cycle completes a full circle. Data mining acts on any or all of the data sets generated during the stages described now.

Stages in the work flow of taxonomic research and services (overview in Figure 1):

Collection making. Collecting-event data include at least a locality name and a date. Further factual data can be anything relevant to the individual naturalia – the data recorded vary according to subdiscipline. Frequently recorded data, in addition to locality and date, may include: the collector's name, site details (for instance, site altitude above sea level, topography, macro-, microhabitat, stratigraphy[3]), collecting technique, associations with other objects (like host organisms), and unique sample codes.

Modern collecting efforts usually make use of technological tools, such as GPS receivers for determining geographic co-ordinates; altimeters for determining altitude; and personal digital assistants (PDAs) for recording data. Serious field campaigns result in detailed collecting reports, including station and sample lists containing these data. The data is stored in standardized digital form, so that printed labels can be generated and directly attached to the single- or multi-specimen objects collected in the field — for instance, a sample of spiders swept from the vegetation. Extant collections, not coming directly from strategically planned work, may account for up to more than 90% of a museum's collection. Collecting strategies, where applicable, may have been highly biased, being, for instance, concentrated on rare or conspicuous objects, on particular groups of organisms, or on particular themes, rather than on a random sample of the real world.

Collection storage management. Collection storage management includes operational item inventory and registration. After initial preservation, provisional sorting and identification, further conservation and systematically correct curation of the naturalia in the collections is the next stage. The physical arrangement may be based, as much as possible, on the Tree of Life mentioned before. In practice this means a quasi-hierarchical or linear arrangement as published in a recent digital or hardcopy species catalogue. The registration of accessions and their physical position on the collection shelves, with or without the data from the previous stage, takes place during this in-house stage (2) in the workflow. Collection registration software is applied extensively, sometimes integrated with functions relevant to the preceding and subsequent research stages. Modern multidisciplinary registration software (including heavy systems such as BioLink and Specify) handles hundreds of data types and includes taxa cataloguing and geographic tools, as well as additional scientific functionality. Simple bulk data import/migration modules may simplify registration tasks. In this stage again, implicit and explicit selection strategies and individual interests may provide filters for the actual inclusion of material in the collection, and data quality as such may be a moot point [Wilkie, 1999]. Therefore subsequent data mining should proceed circumspectly.

3 Geological: soil layer data.

Pattern recognition. Taxonomic research improves identification, curation, and data management. Integral to the responsibilities of the curator is taxonomic and nomenclatural research on museum specimens, the product of which feeds back into the shared knowledge of identification and classification, thus forming a feedback loop to better improve the initial steps of museum curation. The taxonomist usually inspects many different collections from different institutions, verifying and refining identification guides, analyzing the numerous physical attributes (characters and character states) of the objects in the material under study. Comparisons, both with published results from colleagues (starting with Linnaeus) and with specimens in present day working collections, are indispensable. The work of the taxonomist cannot be conducted without access to the publications in the large historical reference libraries maintained by museums. Additional collecting of material in the field to resolve discrepancies or to fill gaps in the information spectrum may be required. Operationally recognizable taxonomic units (abbreviated OTUs) are distinguished, and each is then defined by particular character states. New character tables and dichotomous keys to assist in the identification of OTUs are set up. Software can be used to automate this task, some based on clustering algorithms leading to dichotomous keys (e.g. DELTA and similar systems, [Pankhurst, 1991; Dallwitz, 2000]). Usually full descriptions and illustrations are produced. Names are assigned to OTUs, usually down to species level, and conforming to the nomenclatural rules. Some species may prove to be new to science (and must be given a new scientific name), some species names will prove to be synonymous with other, previously described names, and many locality records may be new as well. The distribution is mapped. Certain data are ploughed back into the registration database of stage 2, resulting in record increases and updates, both on object and taxon levels. This stage may be iterative.

Process hypotheses. This results in generating trees for evolutionary reconstruction. A natural progression beyond the taxonomic revision is phylogenetic research into the evolutionary relationships of the organisms under study. While many museum curators choose to pursue this task, it is not necessarily an integral part of their duties with regard to the collections. Inferring phylogenetic history involves identifying a set of supposedly homologous characters (be they morphological traits or DNA sequences), and discovering a hierarchical bifurcating tree that best resolves the collective set of shared derived character states. Cladistics and phylogenetics are the terms applied to the science that specializes in this evolutionary tree reconstruction. Discovering, storing, and retrieving phylogenetic data in digital form has recently become a subdiscipline in its own right, particularly with the advent of molecular techniques that have vastly expanded the discipline.

Interdisciplinary enrichment and synthesis. This stage investigates data congruence with process models and mechanisms, including its contextual embed-

ding. The final research stage is the recognition of common mechanisms in evolutionary processes and their environmental context on various spatiotemporal scales. The outcomes of biodiversity work may be at odds with, or be falsified by external so-called para-data, like Earth science data. For instance, land masses facilitating migration and evolution of organisms cannot be postulated where sea levels evidently render this impossible. At this stage, integration of different Life and Earth science disciplines takes shape in order to reconstruct the 'big picture'. What about the ongoing change of biodiversity patterns, their bearing on ecosystems, and vice versa? How are biological and Earth system events jointly shaping the Tree of Life patterns recognized? Can Geographic Information Systems (GIS) elucidate the processes, and reveal trends into the future? What sorts of predictions can one make about the future of Man's biotic environment — his life support system, and how reliable are they? What of the sustainable use of biodiversity in the next decades? Few scientists contribute directly to answers to these 'big questions', but let it be clear: the biodiversity industry as a whole continually contributes indirectly, by both adding and revising data in the vast data warehouse constituted by the global information network of collection facilities.

Meta-data recycling. In this stage thesauri, glossaries, dictionaries, and catalogues are produced. Higher-level content produced in the previous stages, once standardized and accepted, is ploughed back into the system, improving data quality and exchangeability. These may be included in software packages, they may be generated by them, or they may be taken from the web — often in the form of lists of taxonomic names, geographic names, ecological terms, author and collector names, and bibliographic references, etc. The standardization of digital thesauri and pick lists used with recording software will increase the efficiency of future data mining considerably. However, few standards have as yet been globally accepted, although most taxonomists find no problem in an operational acceptance of the higher taxa as used by the major abstracting (meta-database) facilities. The Catalogue of Life [Species 2000, 2001] is a good example of a no-nonsense approach to a standard world index of described species (discussed below).

Spreading the word: actual output, marketing, services, applications.

Ultimately, data should be translated into knowledge geared to various actual and potential end user groups. These range from fellow scientists to policy makers to the general public, the information varying accordingly from technical scientific publications and reports to educational and recreational products, possibly extending to presentations simply raising awareness and interest in the 'big picture'. Museums are ideally positioned to do this. The larger taxonomic facilities, apart from external channels, usually have their own scientific outlets in the form of hardcopy technical publication series, and E-publishing projects are under construction almost everywhere. Long-standing global tertiary-type pub-

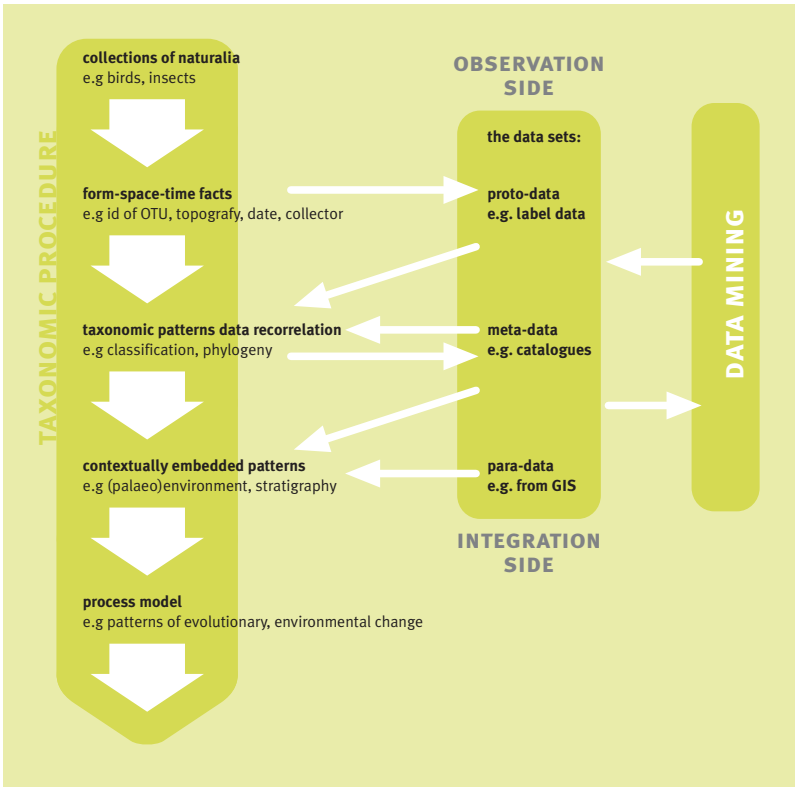


Figure 1
Museum and herbarium collections generate different data sets in different stages of the taxonomic work flow. Central items of study are objects representing units known as taxa (e.g. species). Various categories of meta-data are derived from collection-based proto-data and reused iteratively to arrive at a hierarchic classification, with or without an underlying phylogeny. Para-data from related disciplines are taken into account in reconstructing the 'big picture' of diversity evolution in an environmental context. Data mining acts then on the data sets generated, and typically involves critical recycling and multidimensional recorrelation of extant data.

lications that have functioned, to put it in ICT terms, as data warehouse portals in biology and geology (e.g. Zoological Record and the Index Kewensis) already went digital around two decades ago, and nowadays there is much more (see [Winston, 1999]). Recent ICT developments greatly facilitate the synergistic usage of the global biodiversity data warehouse and the so-called GBIF (pronounced jee-biff: Global Biodiversity Information Facility of the OECD), under construction, is intended to stimulate this synergism. Some look at GBIF as a future data miner's paradise. Societal applications range from biodiversity conservation to agriculture, fisheries, health, forensics, environmental management, and so forth.

Approaches, examples, and tools for data mining

In general, accessing museum information involves two basic approaches: record-oriented questions and pattern-oriented questions [Dworman, 2000]. Knowledge that emerges from each of these approaches relies on different assumptions, depending on the approach. As can be seen in the examples listed below, many questions combine record- and pattern-oriented access. Furthermore, database tools, even if designed largely for record-oriented searches, can be recruited to perform pattern-oriented tasks when used in combination. These issues are outlined in the next paragraphs.

Record-oriented access

Record-oriented data mining consists of querying individual specimens or specimen information in order to recover knowledge. These queries can be fairly simple, such as the computer-based question 'how many butterfly species are recorded for South Holland in the museum's collections?' or the realia-based question 'what was the first recorded instance of AIDS in Great Britain?' Mining by computer on a flat file might use a straightforward SQL SELECT statement.

Pattern-oriented access: data mining

As [Dworman, 2000] describes it, pattern-oriented questions seek to observe the 'forest from the trees'. Instead of querying for a specific selection of records, pattern-oriented searches seek to observe trends in the distribution of different kinds of collected realia or data. The seminal paper by [Trivers and Hare, 1976] exemplifies this approach. These authors were puzzled by the act that haplodiploid eusocial insects did not appear to show a strongly biased ratio in female to male offspring that would be expected under a kin selection theory. They decided to test whether differences in the sum total mass of male and female offspring could account for this discrepancy, and so they set about weighing ants from whole colonies conserved in the Museum of Comparative Zoology at Harvard University. Remarkably, Trivers and Hare demonstrated that the investment in sex-specific biomass followed a 3-to-1 ratio, while slave-making ants and solitary insects showed a more equal investment — exactly as predicted by theory. The point here is that this pattern was only observable collectively; it depends not on individual records, but on the relative distribution of whole collections of records.

For the most part, computer-based data mining has to work with relational database models that are primarily designed for record-oriented searches. In order to apply pattern-oriented queries, [Dworman, 2000] advocates a procedure whereby the relational database is transformed into a 'data cube', or the so-called 'crosstab' form as implemented in Microsoft Access [Balachandran, 1998]. Such multidimensional representations seek to cross-reference one field with another (in a two-dimensional matrix), or with yet another field (in a cube), and so forth for as many dimensions as needed.

This procedure can be illustrated by the example in Inset 1.

Inset 1: Polymorphity demonstration by [Piel, 2001a] for the spider Metepeira tarapaca

He proved that the spider is polymorphic for the amount of black or white on its sternum. The markings on specimens from Harvard University's Museum of Comparative Zoology and the American Museum of Natural History correlated with their collection elevation, suggesting a thermoregulatory function. As a simple two-dimensional data mining problem with continuous data, this problem was easily addressed using a

graph [Piel, 2001, fig. 108]. But one could equally well transform the data into a crosstab matrix. Seeing this is a simple example, one needs only two-dimensions, but given that the data are continuous, they must be divided into bins — in this case three bins are chosen for each variable (See Table 1).

Table 1

		elevation (m)			
		0-1,500	1,501-3,000	3,001-4,501	
	0-33	0	0	2	2
white line (%)	34-66	0	3	8	11
	67-100	6	4	2	12
		6	7	12	

A crosstab of 24 female Metepeira tarapaca specimen data in three bins of elevation data (columns) set against three bins of sternum coloration (rows). A total of twenty-five observations (n) are distributed among the cells in the matrix.
Prediction analysis makes use of a crosstab observation matrix in concert with a hypothesis error matrix (Table 2) to estimate both hypothesis value ('dell' ∇) and the expected error rate for a prediction, known as the 'precision' U [Balachandran, 1998; Hildebrand, 1977a; Hildebrand, 1977b]. Hypotheses (as expressed by the error-cell matrix), the size of the bins, and the dimensions of the data cube, can all be manipulated and massaged so as to maximize ∇ and U jointly. This manipulation and data exploration is, in essence, knowledge discovery and data mining. As described in [Balachandran, 1998] hypothesis and precision values are calculated as follows:

Formula 1

$$\nabla = 1 - \frac{\sum_{i=1}^{R}\sum_{j=1}^{C} \omega_{ij}\ n_{ij}}{\sum_{i=1}^{R}\sum_{j=1}^{C} \frac{\omega_{ij}\ n_{i\bullet}\ n_{\bullet j}}{n}}$$

Formula 2

$$U = \sum_{i=1}^{R}\sum_{j=1}^{C} \frac{\omega_{ij}\ n_{i\bullet}\ n_{\bullet j}}{n \cdot n}$$

where:

R = number of rows

C = number of columns

n = total number of observations

$n_{i\bullet}$ = number of observations in category i for the row variable

$n_{\bullet j}$ = number of observations in category j for the column variable

n_{ij} = number of observations in cell ij

ω_{ij} = error value in cell ij

In the Metepeira tarapaca example, the hypothesis value, ∇ = 0,98, is very high, while the precision value, U = 0,56, is reasonable, indicating that the high hypothesis value does not result from a powerless error cell matrix (such as one consisting entirely of zeros, in which case U would indicate zero). Note that this example is very simple — complex hypotheses would involve a data cube with many more than two dimensions.

Table 2

		elevation (m)			
		0-1,500	1,501-3,000	3,001-4,501	
	0-33	1	1	0	2
white line (%)	34-66	1	0	0	1
	67-100	0	0	1	1
		2	1	1	

An error-cell representation of the hypothesis that the coloration on the sternum of Metepeira tarapaca has a thermoregulatory function. Spiders at low altitudes, where the air temperature is high, are expected to have more white on their sterna to shield them from excessive heat. Spiders at high altitudes, where the air is cooler, need not reflect so much sun, and do better by showing more black coloration to capture the sun's warmth. Each cell where an observation is expected is given a score of zero; each cell where an observation is not expected, is given a score of one.

Researchers should always be aware that the results of museum data mining, and in particular pattern-oriented methods, must be scrutinized for confounding effects, especially those that result from biases in the acquisition of realia. Generally speaking, conclusions will usually be of a more qualitative rather than a quantitative nature, because realia are usually gathered through collecting methods instead of the less biased sampling methods [Scoble, 2000]. Obviously, any statistical conclusions of a quantitative nature require that the sampled data be a fair representation of the target population.

Examples of museum data mining

Much has been discovered from data mining natural history museums, and this is a harbinger for many future discoveries [Brooke, 2000]. Data mining of biodiversity collections is usually performed on a data mix of uncritical museum data, critical[4] museum data, non-museum-based data sets (e.g. climatic and geographic data), and information from the literature. Collections made with predetermined research objectives only come under the current definition of data mining when these objectives change. Such collections may, for instance, be made in the context of a faunal survey, and analyzed for the production of distribution maps. Yet, the preservation of morphological and anatomical features

4 Assembled expressly with specific research objectives, questions, or hypotheses in mind.

recorded along with spatiotemporal data may, for instance, enable the reconstruction of life cycles or show geographic patterns of morphological change. Knowledge extraction becomes ever more powerful, when uncritical and critical museum data are 'value added' by cross-referencing these data with authority files and other non-museum-based data sets. For example, simple cross-referencing locality information with gazetteer databases (e.g. NIMA name server[5]) allows for GIS analysis using mapping software; cross-referencing co-ordinate points with hypsographic databases (e.g. at Harvard[6]) allows for analysis with an altitudinal component. Linking a multitude of separate data sources in this way in order to glean emergent knowledge is the ultimate goal of meta-analysis projects such as GBIF.

Broadly speaking, such meta-analyses can be divided into several types: those having to do with taxonomic inventories and ecological distribution, in which taxonomy and geographic co-ordinates are key, and those having to do with time series, in which specimen dates are key [Brooke, 2000]. For time series analysis Hidden Markov Models (see 6.2.10) are an often used method.

Species numbers and extrapolation

Questions of species richness and related policy issues are high on the biodiversity research agenda (see Introduction). The vast majority of organisms cannot be counted and assessed without data mining both old and newly made collections, mining existing sets of meta-data, and doing subsequent in-depth taxonomic research. The debate on how many species of organisms exist on Earth was re-kindled following the estimates made by Terry Erwin of the Smithsonian [Erwin, 1982]. The principle he employed for estimating species numbers involves extrapolating the diversity of microhabitats, such as the insect fauna associated with tropical tree species, to the level of the entire biome. A phytophagous[7] guild of insects may be wholly dependent on a single tree species, and taking into account the fact that not all insects are monophagous[8] and not all trees have equivalent guilds, multiplying the number of insect species with a certain number of tree species results in global and local species number estimates. Measurements of species-specificity are obtained by comparing different microhabitats (such as different tree species). Erwin primarily worked with beetles, a group which accounts for a third of the known total of animal species, i.e. approximately 350,000 — the current most conservative estimates of the actual numbers of beetle species on earth run into tenfold figures. The accuracy of figures from 'Erwinoid' inventories increases with the volume and quality of local inventorying, monographic research, and subsequent re-analysis, including data recycling and data mining. Local sampling, with or without global extensions, may result in remarkable data, useful as reference in biodiversity monitoring on various geographic and ecological scales (examples in [Gleich, 2000]).

5 http://www.nima.mil/gns/html/

6 http://www.herbaria.harvard.edu/~piel/find.html

7 Plant eating.

8 Eating from one source.

Ecological distribution

Taxonomic revisions serve as the primary published scientific contribution of museum staff and our most fundamental knowledge of biodiversity. Many taxonomic revisions focus on applying correct nomenclature and describing the morphology of species. However, with a bit of data mining, the taxonomist can glean some valuable information about the distributional ecology of the species, even when working with exclusively dead museum specimen. For example, [Piel, 2001a] reported a number of ecological observations derived only from museum collection locality data. In the case of the spiders Metepeira vigilax and M. rectangula, localities were cross-referenced with geographic co-ordinates using the NIMA gazetteer server [9], and elevation was estimated for each geographic co-ordinate by cross-referencing with a hypsographic database [10]. The result was a clear example of biotope-specific habitat preference in which the elevation of species localities decreased with degrees from the equator, yet both species showed separate habitat-specific clines (Figure 2).

The date in the year when a specimen was collected provides information about the seasonal occurrence of the species. However, seasonal occurrence can vary with elevation and latitude, so merely reporting the time of year, when mature specimen were collected is not necessarily very informative as to when and where a species can be found. In the case of Metepeira glomerabilis, the availability of mature specimen shifts with elevation: at low elevations the spiders mature earlier than at higher elevations (Figure 2).

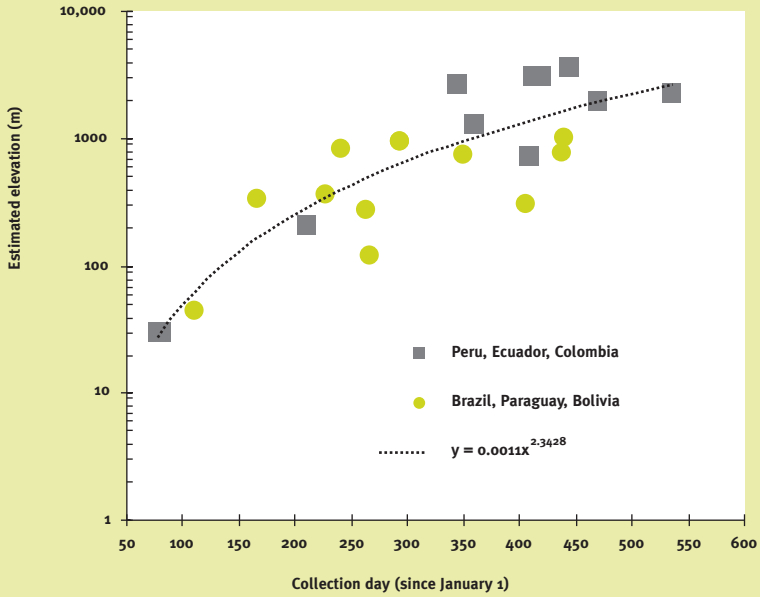


Figure 2
Elevation at which mature Metepeira glomerabilis were found as a function of collection day. The time that it takes for this species to mature appears to increase with elevation, with higher altitude specimen taking much longer to mature. This relationship suggests the possibility of an annual cycle for spiders at the lowest elevations and a biannual cycle for spiders at the highest elevations, with intermediate elevations showing a broad seasonal distribution. Judging by the scattering of the grey squares (Peru, Ecuador, Columbia) and green circles (Brazil, Paraguay, Bolivia), this relationship holds regardless of whether the specimen were collected on the eastern side or the western side of the Andes. Redrawn from [Piel, 2001a].

9 http://www.nima.mil/gns/html/

10 http://www.herbaria.harvard.edu/~piel/find.html

Similarly, Metepeira galatheae that live farther from the equator have a narrower season than those living in more tropical areas (Figure 3). Of course, these examples assume that collecting efforts made by biologists are fairly evenly distributed throughout the year. Clearly, academic research seasons, collecting seasons, teaching periods, vacation times, and inclement climate can interfere with this assumption.

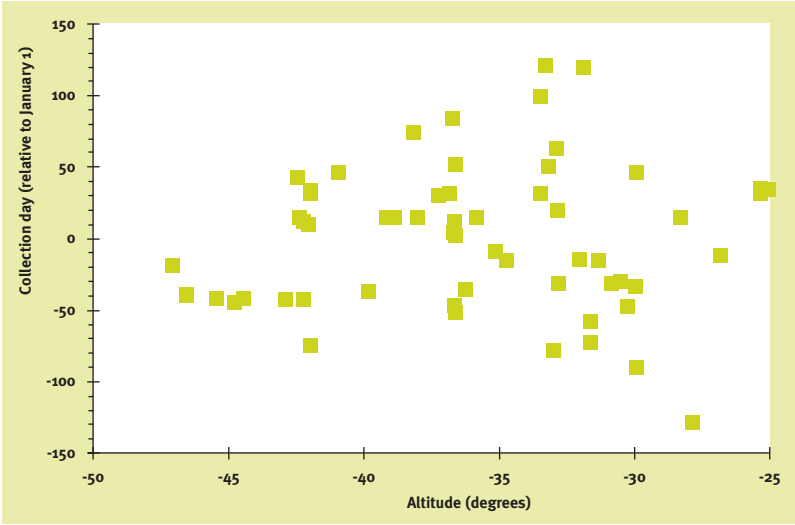


Figure 3
Seasonal occurrence of mature Metepeira galatheae relative to altitude. Mature spiders at temperate altitudes, such as 45° to 50° south, have only been found during a brief period in late November and early December. In contrast, at more tropical altitudes, such as 35° to 25° south, mature specimens can be collected over a much broader period of up to 260 days out of the year. Redrawn from [Piel, 2001a].

Time series knowledge: historical time

The fact that the collection date is usually recorded for each specimen in a museum is valuable, because it provides a historical time series that scientists can analyze retrospectively. This type of analysis can be important in assessing environmental pathology, such as the concurrence of physical changes in specimen with the introduction of certain toxins in the environment. Historical changes in the attributes of species may be extracted from collections, a well-known example being eggshell thinning in birds and melanism in moths. [Ratcliffe, 1970] and others showed that under the influence of the insecticide DDT in birds of prey, eggshells become thinner and the reproductive success of the birds is accordingly reduced. In another case [Green, 1998] examined museum collections of thrush eggs to show a trend towards thinner eggs starting in the nineteenth century. Based on the emergence of this trend, Green was able to rule out the pesticide DDT as the causative agent. Environmental pathology can also impact the population genetics of a species. A textbook example of this phenomenon is industrial melanism. During the 19th century populations of the Peppered moth (Biston betularia) became, as was apparent from their frequency in historical collections, more blackish in industrial areas where air pollution was high. This phenomenon was explained by differential predation pressure from birds. In industrial areas pollution eliminated lichen growth on the trees

where the moths rested during the daytime. Birds then preyed more heavily on the uncamouflaged lichen-colored variety, whereas in non-polluted regions it was the melanic mutants that were more easily spotted on the lichen-clad tree trunks and branches (for a review see [Majerus, 1998]).

Time series can also shed light on epidemiology, since many human diseases have pathogens that use other animals as vectors and reservoirs. [Persing, 1990] analyzed museum collections of the tick Ixodes dammini in order to determine the earliest occurrence of Lyme disease in North America. Since the ticks are the main vector of the Lyme disease pathogen, Borrelia burgdorferi, Persing et al. were able to show that the disease was first introduced to New England long before the first human incidence was diagnosed. Similarly, [Marshall, 1994] examined tissues from museum skins of the white-footed mouse, Peromyscus leucopus, in order to show that this animal was serving as a reservoir for the Lyme disease pathogen as early as 1894. Collections are frequently databased to a certain level purely for management purposes (see under collection storage management)), but even then the collecting event data and other proto-data, are usually recorded, also by the numerous non-professional collectors associated with museums and herbaria. Organizations of collectors in some subdisciplines (entomology, malacology) provide their members with sophisticated database tools to streamline their recording activities, so that data mining can be done more efficiently. The study of the micro-moth fauna is an interesting case in point. Recent changes in both species ranges and species seasonality were established by mining the extensive database extracted from collections made largely by non-professional entomologists [Ellis, 1997a; Ellis, 1997b; Ellis, 1999]. The shifts they describe in geographic, seasonal, and diversity-abundance patterns over decades are correlated to higher temperatures and possibly to other climatic factors. Enormous databases similar to the micro-moth collection database are in existence, but data mining projects of any size usually focus on observation databases of vertebrates and flowering plants and not on 'the other 99% of Life on Earth' [Ponder, 1999], where extensive systematic realia sampling (see collection making) usually precedes databasing.

Data from museum collections can offer straightforward information about the meaning of geographic names and localities. In cases where a collector has not accurately specified the location of a collection point, one can only infer the correct location through geographic analysis of the museum collection data as a whole. In the case of Willis Gertsch's Mexican collecting trip, summer of 1947, one locality was listed as only 'La Loma' on August 20th. According to NIMA[11] (there exist more than 60 Mexican locations called La Loma — so which one did Gertsch visit? One solution is to data mine the collection information for all localities collected by Gertsch in the vicinity of August 1947. These records revealed his exact itinerary (Figure 4), and consequently the correct La Loma can be selected based on where he was just prior and just post August 20th.

11
http://www.nima.mil/gns/html/

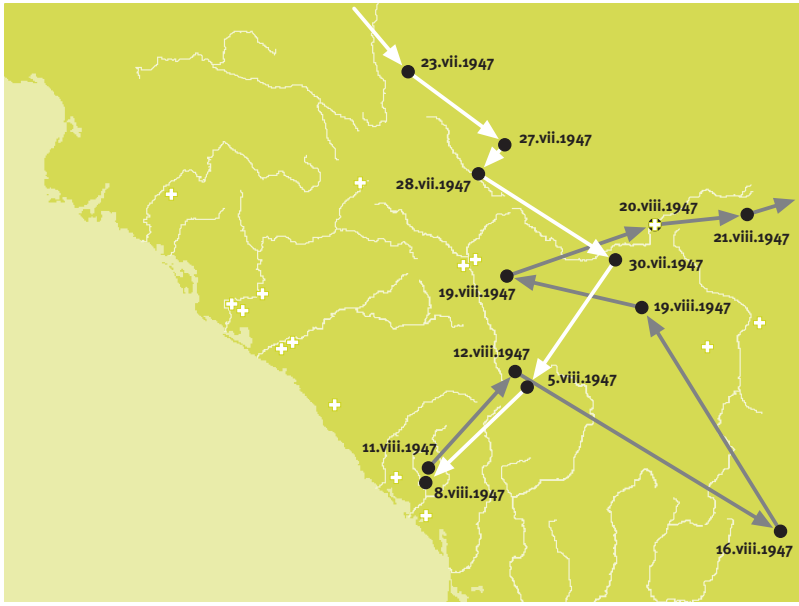


Figure 4
Collecting itinerary of Willis Gertsch in Mexico, July-August 1947. Each circle indicates a point where Gertsch collected Metepeira spiders, based on specimen labels in the American Museum of Natural History. Black crosses are a subset of over 60 Mexican localities that go by the name 'La Loma'. Only through analysis of Gertsch's itinerary can one select the correct La Loma, since only the place name was given on his specimen label dated August 20, 1947, and this alone did not provide enough information to determine the locality.

Time series knowledge: geological time

Fossil collections in museums provide time series information not so much with regard to their collection dates, but more with regard to their geological time of origin. These dates can provide all kinds of knowledge, and even shed light on the effects of global climate change. As fossils have a long distribution in time, they can provide us with information on processes on time scales of millennia and higher, and on phenomena that took place long ago, in 'deep' history [Benton, 1999; Johnson, 1999]. The analysis and subsequent synthesis of changing faunas and floras of fossil sites enables testing of long-standing theories and hypotheses, ranging from the big plate-tectonic issues to details on orogenic, island, volcanic, ocean level, glaciation events, etc. For example [Hellberg, 2001] used museum collections of fossil snail shells to show that Late Pleistocene climate warming caused explosive morphological evolution in a marine gastropod.

The historical biogeography of Southeast Asia is being re-written by recycling and mining old data, and filling knowledge gaps with additional query-driven exploration. An illustrative attempt to reconstruct the Quaternary biogeographic history of the Southeast Asian islands using a numerical analysis of a large ultimately collection-based set of meta-data concerning bats, birds and butterflies, is the paper by [Holloway, 1968], and many have followed since, stretching history, with the inclusion of new tectonic data, back to Mesozoic times.

Valuing and conserving richness and uniqueness

Irrespective of the time dimension, meta-data supplemented by collection- and field-event proto-data, may be mined effectively for conservation purposes. In a

seminal paper on the valuation of bird faunal hotspots on a global scale, i.e. high concentrations of bird species with very limited distributions, the International Council of Bird Protection [ICBP, 1992; Stattersfield, 1998] demonstrated the value of old and new distributional data in locating and prioritizing biodiversity hotspots for conservation. [Vane-Wright, 1991] and others went further by including the evolutionary dimension, characterizing areas in terms of taxon (species) numbers, endemism percentage, and evolutionary history as evident in the classification hierarchy. To put it in extremes: all maintain that there is some sense in giving areas with high species numbers including a high percentage of geographic and phylogenetic species isolates a conservation priority higher than those with low numbers including a high percentage of widespread as well as closely related (very similar) species. The WorldMap program computes scores on a global scale[12], but the principles can of course be applied at different scales anywhere and at any time. Where critical phylogenetic data are not available, the classification hierarchy or other diversity indicators may be introduced as a working hypothesis.

The billions of taxon-related proto-data stored in the large databases of the European Invertebrate Survey (EIS) and similar zoological and botanical equivalents (in The Netherlands united in the VOFF, the Association Flora and Fauna), are used on national and regional levels, again taking into account qualitative and quantitative patterns of taxonomic diversity and local faunal uniqueness. The Dutch EIS database alone contains millions of records, and can indeed play a significant role in analyzing regional diversity patterns, as exemplified by the book series Nederlandse Fauna [e.g. Turin, 2000; over 100,000 records, mostly pre-1960, collection-based]. Similar databases and their concomitant collections exist elsewhere, but the fact is that they are underused and that their data mining potential remains hugely underestimated.

An organization constantly mining and recycling published conservation-relevant data through a complex set of meta-databases is the World Conservation Monitoring Centre[13], now associated with UNEP (United Nations Environment Program). WCMC has for many years mined and monitored data concerning the status of species, ecosystems, and faunas and floras worldwide. Red Lists of threatened species, inventories of threatened ecosystems, country assessments are some of the products and many are now also available on CD-rom or can be accessed via the Web. Any data concerning 'the other 99%' appearing in a conservation context is likely to have something to do with taxonomic facilities and collections.

Data mining tools

How can the information behind the taxonomic tags be found? Old and new data are stored, and made accessible in numerous ways dependent on their target audience. Essential to taxonomy, and to other disciplines tackling biodiver-

12 Vane-Wright originally included bumblebee data to illustrate the use of the program, but there is more (see [Williams, 1992] plus the web site).

13 www.unep-wcmc.org

sity in descriptive (i.e. non-experimental) ways, are the current tertiary sources, including the Zoological Record and the Index Kewensis, and the various abstracting journals (see [Winston, 1999]). These sources have all gone digital, and are increasingly accessible via the Web on the basis of taxonomic, geographic and thematic searches. Taxonomic thesauri are under construction everywhere, an effective example being the Species 2000 Catalogue of Life [Froese, 2001]. By entering a taxonomic name, either scientific or English, either directly or by searching through the hierarchic classification, links are established to records in a large federation of specialised, not necessarily uniform databases on the Web. Entering the English fish name Sole, gets you in a few clicks to the fish genus Solea in FishBase, a global database of all fish species, giving more information on your favorite fish species, as far as available on-line. The Catalogue of Life is a robust, straightforward search engine, and potentially capable of easily querying taxon-based data on different geographic scales in a variety of world-wide, decentralized databases — one could call it a global federation of databases. A rapid expansion of the on-line search options is anticipated, for instance, as the GBIF initiative takes off (see the next section).

Expert systems for identification and diagnosis

Attribute data (characters and character states) of items (taxa, or their individual representatives) extracted from both collections and literature may be converted into strictly defined text formats, which then can be handled by expert system software, with or without multimedia support (pictures, sound). In principle, it is, with the usual voluminous data sets, simply impossible to predict which of the data included will prove to be useful with respect to the primary objectives of the expert system, such as the identification of taxa. A large data matrix consists of hundreds of character state definitions for hundreds of items. A widely used system is exemplified by DELTA, which has many options [Dallwitz, 2000; Pankhurst, 1991], and the algorithm works on the hill-climbing principle of finding the shortest route in a decision tree to the name of a single taxon. The characters are ranked accordingly, the best coming at the top of the list. Selecting the appropriate state of the best characters (such as head color or body length) for the specimen under investigation results in the elimination of unfit items and their characters, followed by the re-ranking of the remaining characters, another selection, and so on, until a single taxon name is left – or none, in which case the taxon is apparently undescribed (not present in the set of items). DELTA type software can not only be used interactively, but written dichotomous keys and diagnoses can be outputted as well.

Non-text data for identification and diagnosis of items

Whereas the aforesaid type of expert system is based primarily on handling text-based and numerical data, it is anticipated that other data types, like

images of organisms, audio data (e.g. bird song), and physicochemical data (e.g. pheromones, smells, etc.), will play a far greater role in future expert systems and underlying databases. Whereas in certain scientific disciplines, for instance in forensic research, optical image recognition plays a prominent role in the comparison of items (e.g. fingerprint or eye iris biometrics), this is still an underdeveloped area in biodiversity collection facilities. The identification of a bird using an image and or audio recognition system linked with an adequate database can easily be envisioned. In practice, however, the development of these types of sophisticated expert systems has scarcely begun. Capturing images, sounds, and smells for automatic identification, and sending bird watchers the identification of a bird via their portable network-linked system as soon as they focus in on a particular specimen, seems to be a far-cry from present-day multimedia CD-rom's. For the time being most image identification and retrieval action in museums still requires the support of human experts, despite the fact that such experts are rapidly becoming rare.

Phyloinformatics and the Tree of Life

Reconstruction of phylogenetic history often makes good use of museum specimens. Evolutionary history can be reconstructed using morphological characters obtained by examining museum material, as well as by sequencing DNA from organic fragments of preserved museum specimens. Although DNA work is most easily performed on relatively freshly collected tissues, careful handling methods and protocols for ancient DNA allow sequencing from fossils [e.g. Sorenson, 1999] and older museum material — even for species that have since gone extinct [e.g. Cooper, 1994].

In addition to reconstructing evolutionary history, phylogenics also provide a valuable tool for comparative analysis: not only do they summarize character homologies in the most efficient way possible, but they also have a high degree of predictive value for patterns of poorly known character homologies. For example, models for oxygen isotope exchange in terrestrial plants are used in studies on global warming, and since C_4 plants and non-C_4 plants handle oxygen isotopes in different ways, the relative proportion of such plants are necessary for these calculations. However, [Donoghue, 2001] has suggested that such calculations may need correcting for the fact that C_4 plants do not form a monophyletic group, and that a more inclusive clade that unites all C_4 plants also includes many other non-C_4 taxa. Phylogenetic theory predicts that these non-C_4 taxa actually have a photosynthetic chemistry that is more like other C_4 plants, thus amending estimates of the relative proportion of plants presumed to have one chemistry over another. The upshot is that the phylogeny of green plants can be used to correct models of oxygen isotope balance, and ultimately affect estimates of global warming.

Assembling the Tree of Life is one of the ultimate goals of phylogenetic system-

14 http://www.treebase.org

15 http://www.tolweb.org

atics. Each tree that emerges from analyses of smaller subsets of taxa can be used as the building blocks for constructing a super tree of all life on earth. TreeBASE[14] is a database that stores phylogenetic knowledge in terms of individual trees, while Tree of Life[15] is a database that stores a synthetic super tree of life. With an exponentially growing number of published phylogenies [Sanderson, 1993], data mining techniques and phylogenetic synthesis techniques are needed in order to build the Tree of Life [Piel, 2001b; Sanderson, 1998]. These databases and data mining techniques form the core of Phyloinformatics — a new and growing subdiscipline in Biological Informatics.

International E-networking

Developing and organizing biodiversity databases worldwide

Although many smaller institutions have digitized their collection registration to some extent, none of the larger institutions, i.e. those with 10 million objects and upward, have completely done so, and at best they have a complete registration index of taxonomic names with approximate shelf positions. Clearly there is an enormous amount of proto-data waiting to be processed via comprehensive multidisciplinary software packages. A new organization to promote these efforts is the Global Biodiversity Information Facility (GBIF), which had its origins in the OECD Megascience Forum Working Group on Biological Informatics. The Secretariat of GBIF, which will soon be established in Copenhagen, faces the challenge of stimulating the retrospective entry of the data associated with biodiversity collections, and of creating order and accessibility in the mushrooming mass of database initiatives worldwide. Data mining will undoubtedly become an integral part of this initiative. The relatively small Species 2000 organization, already mentioned earlier, is an example of what can be achieved in a few years' time by gathering meta-data from a number of willing world experts. Regional and local knowledge stored in collections and databases may be tied to a hierarchically organized biodiversity network under the auspices of GBIF. Certain organizations will continue to pursue particular objectives, thematic, taxonomic, regional, or otherwise like the WCMC with its emphasis on conservation, FishBase aimed at ichthyologists, and the European Invertebrate Survey with its emphasis on invertebrate species mapping. The structuring and interconnection of all these initiatives will prove to be a formidable task, but the data mining rewards may well be considerable.

Conclusions

The world's biodiversity collection facilities store a wealth of un-databased and consequently un-minable information. Only a fraction of the proto-data have been properly databased. The retrospective recording of these data from collections, the taxonomic processing required, and the subsequent structured, sus-

tainable and accessible storage of both proto- and meta-data are critical in the facilitation of future data mining operations. Collection-based historical data are crucial in resolving time series dimensions on different scales. The question of the global taxonomic capacity required for data warehouse expansion during the next decades demands serious attention from politicians and administrators. Data mining may result in the generation of economically and intellectually significant patterns of biodiversity in space and time, including predictions on future developments. Detecting changes in the patterns of Life on Earth as a function of environment and time on different scales is the core of biodiversity data mining. For 99% of the world's organisms collections as managed by museums, herbaria, etc. remain the primary source of information on these ingredients of Life on Earth, and thus of the actual life support system of Mankind. With the advent of smaller and cheaper systems database initiatives are mushrooming the world over from institutional down to individual levels, and it is high time that a global network was organized and strengthened in order to make full use of its potential. GBIF is expected to remedy the current chaotic situation. Not only professional taxonomists but also non-professional collectors (including hobbyists, ecologists and conservationists) develop databases. In some subdisciplines more than 70% of the collection volume and the associated unpublished and published data are the product of non-professionals, and capturing these data, by embracing the non-professionals and their institutions, requires a special effort. The extant meta-database infrastructure constituted by the tertiary abstracting services should be incorporated in new initiatives. Apart from the common statistical tools, specific tools for mining biodiversity databases for interesting patterns are still limited in number and find limited application. This has to do with the relatively limited size of the databases, with the non-profit use of the databases, and with the specialist nature of the questions asked. With the rapid development of bioinformatics and the anticipated growth of biodiversity data warehouses, however, this may and should change in order to make effective use of the data. Sophisticated, fast, user-friendly search engines incorporating appropriate statistics, both text- and image-based, tapping a federation of not necessarily uniformly structured data warehouses, may form a solution for the biodiversity-information-hungry world. Developing such an engine and making the system portable may indeed become a special megascience project within the GBIF initiative.

Acknowledgements

Although data mining as such is not (yet) on every taxonomist's mind, several colleagues recognized it to be a core issue in our discipline and kindly offered suggestions and advice, but we of course remain responsible for the present text. Particular thanks are due to Prof A. Minelli (University of Padova) and Dr W.N. Ellis (University of Amsterdam).

References

- Balachandran, K., et al. (1998). MOTC: An Aid to Multidimensional Hypothesis Generation. In: Nunamaker and Sprague. (eds.). Proceedings Thirty-First Hawaii International Conference on System Sciences. IEEE Computer Press
- Baxevanis, A.D., B.F.F. Ouellette. (eds.). (1998). Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. Wiley, New York
- Benton, M.J. (1999). The History of Life: Large Databases in Palaeontology. Numerical Palaeobiology. pp249-283
- Brooke, M.D.L. (2000). Why Museums Matter. Trends Ecology Evolution **15**:136-137
- Cooper, A. (1994). Ancient DNA-Sequences Reveal Unsuspected Phylogenetic-Relationships within New Zealand Wrens (Acanthisittidae). Experientia **50** (6):558-563
- Dallwitz, M.J., et al. (2000). Principles of Interactive Keys. http://biodiversity.bio.uno.edu/delta/
- Donoghue, M.J. (2001). Why the Tree of Life Matters to Global Change, and Vice Versa. Conference on Challenges of a Changing Earth. RAI Conference Center, Amsterdam, The Netherlands
- Dworman, G.O., S.O. Kimbrough, C. Patch. (2000). On Pattern-Directed Search of Archives and Collections. Journal of the American Society for Information Science **51**:14-23
- Ellis, W.N., et al. (1997a). Recent Shifts in Phenology of Microlepidoptera, Related to Climatic Change. Ent. Ber. Amsterdam **57**:66-72
- Ellis, W.N., et al. (1997b). Recent Shifts in Distribution of Microlepidoptera in the Netherlands. Ent. Ber. Amsterdam **57**:119-125
- Ellis, W.N., et al. (1999). Is the Microlepidoptera fauna of The Netherlands shifting northwards? Ent. Ber. Amsterdam **59**: 161-168
- Erwin, T. (1982). Tropical Forests: their Richness in Coleoptera and Other Arthropod Species. Coleopterist' Bulletin **36**:74-75
- Frawley, W., et al. (1992). Knowledge Discovery in Databases: an Overview. AI Magazine. pp213-228
- Froese, R., F.A. Bisby. (eds.). (2001). Species 2000 Catalogue of Life: Indexing the World's Known Species. Software Los Baños. Species 2000, Philippines
- GBIF. (2001). Global Biodiversity Information Facility. http://www.gbif.org/frames/
- Gleich, M., et al. (2000). Life Counts: Eine globale Bilanz des Lebens. Berlin Verlag, Berlin
- Green, R.E. (1998). Long-Term Decline in the Thickness of Eggshells of Thrushes, Turdus spp., in Britain. Proceedings Royal Society, London. Series B-Biological Sciences **265**:679-684

- Groombridge, B. (1992). Global Biodiversity: Status of the Earth's Living Resources. Chapman & Hall, London
- Hancock, P.L., B.J. Skinner. (eds.). (2000). The Oxford Companion to the Earth. Oxford University Press, Cambridge, UK
- Hellberg, M.E., D.P. Balch, K. Roy. (2001). Climate-Driven Range Expansion and Morphological Evolution in a Marine Gastropod. Science **292**:1707-1710
- Herrmann, B., S. Hummel. (1994). Ancient DNA: Recovery and Analysis of Genetic Material from Paleontological, Archaeological, Museum, Medical, and Forensic Specimens. Springer Verlag, New York
- Hildebrand, D.K., J.D. Laing, H. Rosenthal. (1977a). Analysis of Ordinal Data. Quantitative Applications in the Social Sciences, **8**. Sage Publications, Newbury Park
- Hildebrand, D.K., J.D. Laing, H. Rosenthal. (1977b). Prediction Analysis of Cross Classifications. Wiley, New York
- Hofreiter, M., et al. (2001). Ancient DNA. Nature Reviews Genetics **2**:353-359
- Holloway, J.D., N. Jardine. (1968). Two Approaches to Zoogeography: a Study Based on the Distributions of Butterflies, Birds and Bats in the Indo-Australian Area. Proceedings Linn. Society, London **179**:153-188
- ICBP. (1992). Putting Biodiversity on the Map: Priority Areas for Global Conservation. International Council for Bird Preservation, Cambridge, UK
- Johnson, K.G., T. McCormick. (1999). The Quantitative Description of Biotic Change Using Palaeontological Databases. Numerical Palaeobiology, pp225-247
- Koerner, L. (1999). Linnaeus: Nature and Nation. Harvard University Press, Cambridge, Mass.
- Kress, W.J., P. DePriest. (2001). What's in a PhyloCode Name? Science **291**:52
- Levin, S.A. (ed.). (2001). Encyclopedia of Biodiversity. **1-5**. Academic Press, San Diego
- Linnaeus, C. (1753). Species Plantarum. Stockholm
- Linnaeus, C. (1758). Systema Naturae. Ed X. Salvii, Stockholm
- Majerus, M.E.N. (1998). Melanism: Evolution in Action. Oxford University Press, Oxford
- Marshall, W.F., et al. (1994). Detection of Borrelia burgdorferi DNA in Museum Specimens of Peromyscus leucopus. Journal of Infect. Diseases **170**:1027-1032
- Minelli, A. (1993). Biological Systematics: The State of the Art. Chapman & Hall, London
- Minelli, A. (1999). The Names of Animals. Trends Ecology Evolution **14**: 462-463
- Nieukerken, E.J. van, A.J. van Loon. (1995). Biodiversiteit in Nederland. Nationaal Natuurhistorisch Museum/KNNV Uitgeverij, Leiden (in Dutch, English summary)

- NRC. (1999). Perspectives on Biodiversity: Valuing its Role in an Everchanging World. National Academy Press, Washington DC
- Pankhurst, R.J. (1991). Practical Taxonomic Computing. Cambridge University Press, Cambridge, UK
- Pennisi, E. (2001). Linnaeus's Last Stand? Science **291**:2304-2307
- Persing, D.H., et al. (1990). Detection of Borrelia burgdorferi DNA in Museum Specimens of Ixodes Dammini Ticks. Science **249**:1420-1423
- Piel, W.H. (2001a). The Systematics of Neotropical Orb-weaving Spiders in the Genus Metepeira (Araneae: Araneidae). Bull. Mus. Comp. Zool. **157**:1-99
- Piel, W.H., M.J. Donoghue, M.J. Sanderson. (2001b). TreeBASE: a Database of Phylogenetic Information. 2nd International Workshop of Species 2000. Tsukuba, Japan
- Ponder, W. (1999). Using Museum Collection Data to Assist in Biodiversity Assessment. In: W. Ponder, D. Lunney. (eds.). The Other 99%. pp253-256
- Ponder, W., D. Lunney. (eds.). (1999). The Other 99%. The Conservation and Biodiversity of Invertebrates. Royal Zoological Society of New South Wales, Mosman, NSW
- Ratcliffe, D.A. (1970). Changes Attributable to Pesticides in Egg Breakage Frequency and Eggshell Thickness in Some British Birds. Journal of Applied Ecology **17**:67-107
- Sanderson, M.J., A. Purvis, C. Henze. (1998). Phylogenetic Super Trees: Assembling the Trees of Life. Trends Ecology Evolution **13**:105-109
- Sanderson, M.J., et al. (1993). The Growth of Phylognetic Information, and the Need for a Phylogenetic Data-Base. Syst. Biol. **42**
- Scoble, M.J. (2000). Costs and Benefits of Web Access to Museum Data. Trends Ecology Evolution **15**:374
- Sorenson, M.D., et al. (1999). Relationships of the Extinct Moa-Nalos, Flightless Hawaiian Waterfowl, Based on Ancient DNA. Proceedings of the Royal Society of London. Series B-Biological Sciences **266** (1434):2187-2193
- Stattersfield, A.J, et al. (1998). Endemic Bird Areas of the World: Priorities for Bird Conservation. BirdLife Conservation Series 7
- Trivers, R.L., H. Hare. (1976). Haplodiploidy and the Evolution of the Social Insects. Science **191**:249-236
- Turin, H. (2000). De Nederlandse loopkevers: verspreiding en oecologie (Coleoptera: Carabidae). Nederlandse Fauna 3. (In Dutch, English summary)
- Vane-Wright, R.I, et al. (1991). What to Protect – Systematics and the Agony of Choice. Biol. Conservation **5**:235-254
- Wilkie, L., G. Cassis, M. Gray. (1999). Quality Control in Invertebrate Biodiversity Data Compilations. In: W. Ponder, D. Lunney. (eds.). The Other 99%. pp147-153
- Williams, P.H. (1992). WorldMap. Software. Natural History Museum, London. http://www.nhm.ac.uk/science/projects/worldmap/

– Wilson, E.O. (1992). The Diversity of Life. Harvard University Press, Cambridge, Mass.
– Winston, J.E. (1999). Describing Species: Practical Taxonomic Procedure for Biologists. Columbia University Press, New York

Links and further references

See CD-rom version of this article

organization	URL	synopsis
Entrez/Genbank	http://www.ncbi.nlm.nih.gov	National Center for Biotechnology Information
FishBase	http://www.fishbase.org	Database of fish species
IPNI	http://www.ipni.org	Index of plant names
MCZ Entomology	http://mcz-28168.oeb.harvard.edu	Database of insect holotypes at the MCZ, including about 7,000 digital images
NIMA	http://www.nima.mil	Digital gazetteer of the world
Species 2000	http://www.species2000.org	Index of the world's known species
Tree of Life	http://www.tolweb.org	Phylogenetic summary of life on earth
TreeBASE	http://www.treebase.org	Database of phylogenetic knowledge
World Spider Catalog	http://research.amnh.org/entomology/spiders/catalog81-87/	Catalogue of spider names, synonyms, and transfers
UNEP-WCMC	http://www.unep-wcmc.org	Databases for conservation purposes
OECD-GBIF	http://www.gbif.org	Global structuring of biodiversity information
BIOSIS	http://york.biosis.org/index.htm	Zoological Record home page

Table 3
Some web resources related to data mining (for more see [Winston, 1999]).

2.3.6 Data mining in economic science

Ad Feelders[1]

"Let neither measurement without theory
Nor theory without measurement dominate
Your mind but rather contemplate
A two-way interaction between the two
Which will your thought processes stimulate
To attain syntheses beyond a rational expectation!"
Arnold Zellner [Zellner, 1996]

Introduction

Data mining is commonly defined as the computer-assisted search for interesting patterns and relations in large databases. It is a relatively young area of research that builds on the older disciplines of statistics, databases, artificial intelligence (machine learning) and data visualization. The emergence of data mining is often explained by the ever increasing size of databases together with the availability of computing power and algorithms to analyze them. Data mining is usually considered to be a form of *secondary* data analysis. This means that it is often performed on data collected and stored for a purpose other than analysis, usually for administrative purposes.
In this Chapter we consider the possibilities of applying data mining in economic science. In doing so, we must naturally be aware of the considerable amount of research that has already been done in economic data analysis. To what extent can data mining contribute to the analysis of economic data? In answering this question we could consider data mining as a collection of techniques and algorithms that have been developed in this area of research. In doing so we could compare data mining algorithms to data analysis techniques more commonly used in economics, and see if they allow us to answer different questions, or to answer existing questions in a better way. Alternatively, we could also consider data mining as a highly exploratory form of data analysis that is *data driven* rather than *theory driven*. The latter aspect of data mining is most important in this contribution.

This Section is organized as follows. In the next Paragraph we give a brief description of the object of study of economics. Then we will consider economic modeling as a way to apply economic theory to particular problems and as a tool to deduce the consequences of particular assumptions. In order to give empirical content to economic models data is required. We will give an overview of the types of data typically available in economics. After that, we will briefly

1 Dr A.J. Feelders, ad@cs.uu.nl, Utrecht University, Institute of Information & Computing Sciences Utrecht, The netherlands

explain how economic data are used to quantify economic models. Data mining approaches to the analysis of economic data are discussed thereafter. In this paragraph we also consider arguments for and against the use of data mining in the analysis of economic data.
Finally, we summarize the main points of our discussion and give an outlook on possible future developments.

Economic science

Individuals, households, groups, and whole economies can be seen as facing the same problem: they have *objectives* but *limited resources* to achieve them. This limit on resources or constraints forces them to *choose* a course of action in order to achieve their objectives. Economics is the study of such choices: how they are made and what their implications will be. A classic illustration of an economic problem is the situation faced by a consumer. The objective can be thought of as that of obtaining the greatest satisfaction from one's purchases of goods. The constraints are one's income and the prices of goods. If one's income were infinite or all prices zero, then the economic problem would largely disappear. But with a limited income and positive prices one has to choose how, through one's purchases, one can achieve the greatest level of satisfaction.
It is common to divide economics into two main branches, micro- and macroeconomics. Microeconomics deals with the behavior of single, or small units such as the individual consumer, the single firm, and the individual worker. It tries to answer questions such as what determines the price of a particular item, what determines the output of a particular firm or industry, and what determines the amount of hours of labor a particular worker is willing to supply. The defining characteristic of microeconomics is that the unit being analyzed is relatively small.
Macroeconomics is the branch of economics which deals with the behavior of aggregate or average variables, such as total output of the economy, total unemployment, and the average price of all goods produced in the economy. It attempts to explain the behavior of these aggregates or averages and their interrelationships. The defining characteristic is that the unit being analyzed is an aggregate or total.

Economic modeling

In order to apply economic theory to particular problems, and to deduce the consequences of particular assumptions economists often construct mathematical models. Such models typically take the form of (a system of) equations describing the relations between economic variables such as income, consumption, and interest rate. To give a simple example: in macroeconomics it is often assumed that in the short run total consumption C depends on national income Y. We write this as $C = f(Y)$, where f is some function that we leave as yet

unspecified. Other factors, such as interest rate presumably also affect consumption but for the purpose of this analysis we choose to ignore them. This is an example of the pervasive *ceteris paribus*[2] condition often invoked in economic reasoning and modeling. Basically it means that we assume that all other relevant variables (such as interest rate in this case) remain the same. Now to state that consumption depends on income is a rather weak statement. What is meant actually is that when income goes up (down), consumption will *ceteris paribus* go up (down) as well. More specifically it is often assumed that the relation can be described by the linear equation.

Formula 1

$$C = a + b \times Y$$

The economic interpretation of the symbols in this equation is:

C consumption
b marginal tendency to consume ($0 < b < 1$)
Y national income
a autonomous consumption ($a > 0$).

The linear form is often used for convenience or as a first approximation, but is not usually implied by economic theory itself. Economic theory is usually qualitative in nature. It does not for example specify the exact values of a and b in the above equation, although it does constrain a to be positive and b to take a value between 0 and 1. The qualitative nature of such general models is understandable: one would guess that the parameter values would differ from country to country, and that within the same country their values will change over time. Nevertheless, the behavior of a complex economic system may depend crucially on the specific values of these parameters. When economic models are used to support policy decisions, it is usually important to know their approximate values. To give empirical content to qualitative economic theories, statistical techniques are used to estimate the parameters of economic models from data.

Data in economics

In economics almost all available data are of *observational* nature; the data have not been obtained by performing controlled economic experiments, but by passively observing economic reality. One of the consequences of the limited possibility of experimentation in economics is a gap between theory, with its frequent invocation of the *ceteris paribus* clause, and the available data. For example to estimate the demand curve for oranges -the relation between the price of oranges and the quantity demanded — it is not sufficient to observe prices of oranges in different time periods and the corresponding quantities purchased. The reason is that the 'other things' (e.g. income and the prices of other products) have the nasty habit of not remaining equal. In order to make a

2 Latin for (all) other things being equal.

proper estimate of the price-elasticity of oranges we would have to include other important influences on the demand for oranges in our analysis as well. If we were in a position to construct an *experiment* to obtain the relevant data, we could control for those other variables to make sure that the *ceteris paribus* clause is fulfilled.

Having noted that economic data are primarily of an observational nature, we turn our attention to the different types of data structures typically encountered.
Usually the following data structures are distinguished

1 *Cross-section data:* the observation of variables on different units (e.g. people, households, firms). For example, the observation of income of many different people results in cross-section data. The ordering of data does not matter.
2 *Time-series data:* the observation of variables at different points in time. For example, the observation of an individual's income at different points in time yields a time-series.
3 *Panel data:* the observation of variables on different units at several points in time. For example, the observation of income of different people at several points in time results in panel data.

Another useful subdivision of economic data is into micro- and macrodata. Microdata are collected on individual decision making units, such as individuals, households and firms. Macrodata result from aggregating over individuals, households or firms at the local or national level.

Lots of data on economic activity is collected on a routine basis. For macroeconomic data this is usually done by government bodies such as Statistics Netherlands[3] and the Bureau of Economic Analysis[4] and the Bureau of Labor Statistics[5] in the US. A large amount of data is currently available on the Word Wide Web. For example, Statistics Netherlands has an electronic database called StatLine that contains information on many economic and social topics. The Bureau of Economic Analysis and the Bureau of Labor Statistics provide similar services as do government bodies in many other countries. A good place to start the search for economic data on the World Wide Web is Resources for Economists[6] (edited by Bill Goffe).

3 http://www.cbs.nl

4 http://www.bea.doc.gov

5 http://www.stats.bls.gov

6 http://www.rfe.org/data

Econometrics: quantifying economic models

As mentioned before, the relations between variables postulated by economists are usually of a qualitative nature. Consider again the relation between total consumption and national income:

Formula 2

$$C = a + b \times Y$$

With the meaning of the symbols as specified above. In order to give empirical content to such a model, we must have observations that allow us to estimate the unknown parameters a and b of this equation.

The discipline of econometrics concerns itself with the application of tools of statistical inference to the empirical measurement of economic models. Regression analysis is by far the most widely used technique in econometrics. This is no surprise, since economic models are often expressed as (systems of) equations where one economic quantity is determined or explained by one or more other quantities. Note that the economic model we start with is deterministic, i.e. it specifies an exact relationship between consumption and income. When we use observed economic data, for example a time series of consumption and income, we do not expect an exact relationship between the two. Equation (2) would lead to the following econometric model specification:

Formula 3

$$C_t = a + b \times Y_t + e_t$$

Where t is an index for different time periods, and e is the error term. The error term accounts for the many factors that affect consumption but have been omitted from the model. It also accounts for the intrinsic uncertainty in economic activity.

According to the received view empirical economic research should proceed along the following lines (see for example [Hill, 2001], Section 1.6)

1 The process starts with an economic problem or question. On the basis of economic theory we consider what variables are involved in the problem and what is the possible direction of the relationship(s) between them. From this we obtain an initial specification of the model and a list of hypotheses we are interested in.
2 The economic model must be transformed into an appropriate econometric model. One must choose a functional form (e.g. linear) and make assumptions about the nature of the error term.
3 Sample data are obtained, and an appropriate method of econometric analysis is chosen.
4 Estimates of the unknown parameters are made and hypothesis tests are performed, using some statistical software package.
5 Model diagnostics are performed to check the validity of the assumptions concerning relevant explanatory variables, functional form and properties of the error term.
6 The economic consequences of the empirical results are evaluated.

Note the dominant role of economic theory and the modest role of the sample data in this procedure. In the next Paragraph we discuss this issue in more detail.

Data mining in economics

In the previous Paragraph we sketched a picture of economic data analysis largely driven by economic theories and models. In practice economic theory is rarely so detailed that it leads to a unique model specification. There may for example be many rival theories to explain a certain economic phenomenon. Also, the usual *ceteris paribus* clauses of economic theory yield some choices to be made when it comes to the empirical estimation and testing of relationships.

In applied econometrics alternative specifications are often tried, and the specification, estimation and testing steps are iterated a number of times. [Leamer, 1978] gives an excellent exposition of different types of *specification-searches* used in applied work. The search for an adequate specification based on preliminary results has sometimes been called *data mining* within the econometrics community [Leamer 1978; Lovell 1983]. In principle, there is nothing wrong with this approach, its combination, however, with classic testing procedures that do not take into account the amount of searching performed have given data mining a negative connotation. [Spanos, 2000] uses the vivid analogy of shooting at a blank wall and then drawing a bull's eye around the bullet hole: the probability of the shot being in the bull's eye is equal to one. The proper way according to the classical view is to specify the model (i.e. drawing the target) before looking at the data (seeing where the bullet hole is).
Here we discuss two approaches to economic data analysis that part from the classical approach outlined before. They part from this approach in the sense that:

1 The data is used extensively to search for a good model.
2 The models are 'atheoretical' in the sense that received economic theory plays a minor role in the analysis.

These issues are addressed in the next Paragraphs.

General-to-specific modeling

In the introduction, we characterized data mining as the *search* for interesting relations and patterns in databases. We have seen that such a databased search for a good model specification is rejected by the traditional approach in econometrics. In practice, however, researchers would start from their favored theoretical model and 'patch' the model, for example by including additional variables, if the data didn't agree (e.g. if a parameter estimate has the 'wrong' sign). In this procedure one starts with the favored model, which is usually a

relatively simple theoretical model, and repair and extend it to uphold the favored hypothesis, if any data problems are encountered. Different researchers starting from different initial hypotheses will very likely end up with different models at the end of the day.

Others have argued that it is defensible to search the data for a good specification as long as this search is performed in a systematic and justifiable manner. An approach to econometric modeling that explicitly incorporates search is the general-to-specific modeling approach [Hendry, 1982]. The main idea is to start with a complex model and to simplify it through the repeated application of statistical tests on the significance of model parameters. The complex model we start with should ideally include the rival models concerning some economic phenomenon.

The model is taken to be of autoregressive distributed lag (ADL) form.

Formula 4

$$y_t = \sum_{j=0:m} (\beta_j x_{t-j} + \delta_j y_{t-1-j}) + e_t$$

Where m is the maximum number of lags considered. Such models are called autoregressive distributed lag models, because they are a combination of an autoregressive model and a distributed lag model. In an autoregressive model the dependent variable y is explained by its own history (so called lagged values of y). A distributed lag model is a regression model in which the current value of y is explained by current and past values of one or more independent variables x.

[Hoover, 1999] describes a simulation study in which a mechanical search procedure (one could say: data mining algorithm) was formulated , which mimics some aspects of the search procedures used by practitioners of general-to-specific modeling. Full mechanization of the search procedures is very hard, because consistency with economic theory is also used to judge the acceptability of a candidate model. In this simulation study the true 'data-generating process' (i.e. the model that generated the data) is known, and the data mining algorithm is assessed for its ability to recover this true model. They report fairly positive results. And in as far as the algorithm is shown to have some defects (such as a tendency to 'overfit', i.e. inclusion of extra variables in the final specification), they suggest adaptations to overcome these problems. In a reaction to this study, [Hand, 1999] argues that the assumption that the true model is contained in the initial model is not realistic. Therefore one should not measure success by how often the search procedure yields the true model, or a model that includes the true model, but rather by how accurate the predictions of the

final model are. Hand voices the opinion that the structure of the model is irrelevant, because one can never know the true structure, but that models should be judged exclusively on their predictive performance. Needless to say that this is a very controversial point within economics.

VAR models

The discussion concerning the pros and cons of VAR (Vector Auto Regression) models provides a good illustration of the arguments for and against data mining in economics. As mentioned above the 'traditional' approach to learning from economic data relies heavily on economic theory to provide a specification of the model. Economic models typically consist of a number of equations, one for each dependent variable, where each equation describes the relation between the dependent variable and a number of explanatory variables. These models are often referred to as structural models to emphasize that the mathematical equations depict (without exploiting possibilities of algebraic simplification) the detailed economic behavior postulated by the model. Each equation in the structural model either describes a hypothesized pattern of economic response or embodies a definition. A large body of econometrics is concerned with the estimation of such systems of equations on the basis of observed data. In practice it turned out that:

1 Economic theory is usually not specific and detailed enough to arrive at a unique model specification (in other words many different specifications are consistent with economic theory).
2 Because of technical problems with the consistent estimation of such systems of equations, in many cases 'incredible' (from the viewpoint of economic theory) assumptions have to be added to make consistent estimation feasible.

VAR models [Sims, 1980] originated from a discomfort with this situation and also the observation that time-series models (not based on economic theory) were shown to have equal or better predictive performance than the so-called structural models. Essentially VAR models are an extension of time-series models to multiple equation systems. For example, a two variable VAR(p) model looks like this:

Formula 5

$$y_t = \alpha_1 + \gamma_1 t + \sum_{j=1:p} (\beta_{1j} x_{t-j} + \delta_{1j} y_{t-j}) + e_{1t}$$

Formula 6

$$x_t = \alpha_2 + \gamma_2 t + \sum_{j=1:p} (\beta_{2j} x_{t-j} + \delta_{2j} y_{t-j}) + e_{2t}$$

Here p denotes the lag length of the model. Thus if p = 2, we assume that all variables depend on the 2 previous values of all variables in the model, including itself. Of course we may use the data to search for a good value of p.

[Koop, 2000] gives the following example. Macroeconomic theorists have created many sophisticated models for the relationship between interest rates, price level, money supply and real gross domestic product (GDP). A well-known example is the IS-LM model extended for inflation, but there are many others as well.
A VAR modeler would merely assume that interest rates, price levels, money supply and real GDP are related, and that each variable depends on lags of itself and all the other variables. Apart, perhaps, from the variables included in the model, there is no link between the empirical VAR and a theoretical macroeconomic model.

It is interesting to consider the arguments that have been brought to bear for and against VAR models as opposed to structural models. The major argument against their use is that they are not based on economic theory ('atheoretical') and therefore are useless in the advancement of economic science. They may be used for the purpose of prediction of economic variables, but they do not shed any light on existing theory. The major argument in favor of VAR models is that they do not make incredible or arbitrary a priori assumptions concerning the relations between the economic quantities under study.

The basic points of disagreement then seem to be whether prediction per se is a legitimate objective of economic science, and also whether observed data should be used only to shed light on existing theories or also for the purpose of hypothesis-seeking in order to develop new theories. Firstly, in our view prediction of economic phenomena is a legitimate objective of economic science. Models used for prediction may however be hard to *interpret* because they may have little connection with the way we understand economic reality. Secondly, it makes sense to use the data for hypothesis seeking. How do scientists get their ideas for new theories, if not from empirical observation?

Summary and prospects

In this chapter we have considered the role that data mining can play in economic data analysis. According to the 'traditional' view of econometrics, the model to be estimated and the hypotheses to be tested should be specified *a priori* on the basis of economic theory.
The problem that practicing data analysts encounter is that economic theory is seldom specific enough to lead to a unique specification of the econometric model. For example, economic theory is always formulated with *ceteris paribus*

clauses. This seldom specifies the functional form of relations between variables and has little to say about dynamic aspects of economic processes. Because of this practitioners have adopted ad hoc methods of 'data mining'; of using the data at hand to find a good model and testing that model on the same data. Although it makes sense to use the data to find a good model, using the same data for testing violates the assumptions of the testing procedure, unless the amount of searching performed is somehow taken into account.
The pure hypothesis testing framework of economic data analysis should be put aside to give more scope to learning from the data. This closes the empirical cycle from observation to theory to the testing of theories on new data. Thus data mining is not a 'sin', but can be made a valuable part of economic theory construction. A sample of data is mined to find interesting hypotheses, but the test of such hypotheses should be performed on data that was not used to create it in the first place. Of course such a procedure would require that enough data is available. This tends to be a problem in macroeconomic times series. In such cases some middle ground has to be found between complete a priori specification and a purely data based model search. The growing amounts of microdata recorded about individual consumers and their purchasing behavior, however, provide great opportunities for data mining.

References

- Hand, D.J. (1999). Discussion Contribution on Data mining Reconsidered: Encompassing and the General-to-Specific Approach to Specification Search by Hoover and Perez. The Econometrics Journal **2** (2):226-228
- Hendry, D.F., J-F. Richard. (1982). On the Formulation of Empirical Models in Dynamic Econometrics. Journal of Econometrics **20**:3-33
- Hill, R.C., W.E. Griffiths, G.G. Judge. (2001). Undergraduate Econometrics (second edition). Wiley, New York
- Hoover, K.D., S.J. Perez. (1999). Data Mining Reconsidered: Encompassing and the General-to-Specific Approach to Specification Search. The Econometrics Journal **2** (2):167-191
- Koop, G. (2000). Analysis of Economic Data. Wiley, Chichester
- Leamer, E.E. (1978). Specification Searches: Ad Hoc Inference with Nonexperimental Data. Wiley, New York
- Lovell, M.C. (1983). Data Mining. The Review of Economics and Statistics **65** (1):1-12
- Sims, C.A. (1980). Macroeconomics and Reality. Econometrica **48**:1-48
- Zellner, A. (1996). Past, Present and Future of Econometrics. Journal of Statistical Planning and Inference **49**:4-8

2.3.7 Agent systems and emergent behavior in economics and E-business

Han La Poutré [1]

Economics, E-business, and ICT

Modern markets and enterprises are characterized by substantially larger dynamics and diversity than in the past. Among the causes of these phenomena are the progress in information technology and the increasing importance of the creation and manipulation of non-material products (intangibles). Markets thus become increasingly complex systems, characterized by fast changes in customer behavior, the increase of market scale (global internet markets), and alliances of enterprises. Moreover, new and better customer models become available through more advanced data mining techniques. For dynamic markets, it thus becomes important to know whether we can obtain the market behavior from individual customer behavior or from individual behaviors of other market parties, like producers.

In ICT, multi-agent systems [Weiss, 1999] are systems of many software agents that interact with each other. Such systems can be based on cooperating agents designed for a specific common goal or on market-like approaches where agents act for their own interest. In the latter, agents now become autonomous representatives of parties in virtual electronic markets, and there is no overall control over the multi-agent system. So, for multi-agent systems, we can ask similar questions about how the agent system behaves as a whole, as it's behavior emerges from individual software agent behavior.

All the players in these systems are related types of agents: software agents in software systems and economic agents in economic systems (e.g. consumers, producers, and traders in economic markets). We could thus treat them in the same manner. Systems of software agents are actually seen more and more as social systems. This is because such systems are essentially open (extendable with new agents) and very large, and because the behavior of an individual agent can be adaptive to its environment.

Why do we need to understand emergent behavior?

The idea is that we want to know how systems of interacting agents work, based on what we know from the agents themselves. In the case of economic agents, the latter can be some model about customer behavior or other microeconomical properties. In the case of software agents, this can be some description of their behavior based on their software, specifications, or just experiments. For markets, the knowledge of their behavior enables the design and study of marketing strategies or market mechanisms. Similarly for software systems, this knowledge enables the design of architectures, mechanisms, and agent strategies.

1 Prof Dr H. La Poutré, hlp@cwi.nl, CWI, National Research Centre for Mathematics and Computer Science, Amsterdam, The Netherlands, http://www.cwi.nl/~hlp/, Faculty of Technology Management, Eindhoven University of Technology, Eindhoven, The Netherlands.

What is emergence? Processes of change

Emergent behavior of an agent system is the behavior that comes from the simpler activities of its constituent autonomous parts, the agents. These parts (may) adapt their behavior, to interact optimally with other parts. An important instance is an economic market, where the emergent market behavior comes from the behaviors of the players (consumers, producers, traders) in the market. This has given rise to consideration of such systems in an abstract way: how and when can systems, consisting of various interdependent parts, change to be able to execute a certain task [Kaufmann, 1993]. For such abstract systems control by a central party already appears to be very difficult , as in the evolution of animals in nature or the control of markets. Similar observations have been made for very different areas like economical, physical, or neural systems, and very recently for systems of software agents.

How can and could emergent behavior be obtained?

For economical or open multi-agent systems, top-down design and analysis can be very hard. Especially, mathematical analysis of specific adaptive systems is often difficult to achieve [Holland, 1991; Kauffman, 1993]. With, for example, dynamic markets consisting of volatile and autonomous agents, traditional economical models can hardly cope. A more realistic approach is thus desired, in which the self-steering principles of the systems and their constituents have an important place. A way to address this is by an adaptive social simulation, which simulates the learning and adaptivity capabilities in a society (system) of agents through time. This concerns a specific way of computer simulation, in which the social learning aspects are captured: agents can learn strategies and solutions from other agents in the society, as well as by themselves. Examples in the literature are e.g. [Tesfatsion, 2001; Vriend, 2000; Dawid, 1996; Gerding, 2000; Holland, 1991; Bragt, 2001]. In the next Section we describe an elegant example that illustrates some key features of the design and usage of an adaptive social simulation. Afterwards, we conclude with a subsection describing a current learning technique that is suitable for adaptive social simulations.

An interesting example in cooperation

We illustrate the concepts just mentioned with an appealing but simple example, addressing an adaptive social simulation of an interaction game. We describe a central problem of interaction, coming from social sciences: the cooperation game 'Prisoner's Dilemma'. In this game, two agents are in a joint situation and each agent can choose to cooperate or to defect. If both agents cooperate, they will receive a reward pay-off of 3, if both agents defect they will receive the punishment pay-off of 1. If one agent cooperates and the other defects, the cooperator receives the 'suckers' pay-off of 0, and the defector gets the 'temptation' pay-off of 5. Thus, an agent is tempted to defect in case the

other agent cooperates. However, in a society where many encounters between agents take place, cooperation is desirable and rewarding, but may be difficult to achieve.
The latter is better captured in the 'Iterative Prisoner's Dilemma' (IPD), where two agents repeatedly have the choice to cooperate or defect. Then everybody is obviously best-off by always cooperating, since if one agent is defecting at one moment, the other will probably retaliate later. At best, they will receive an average pay-off of 2,5 (if each agent alternately cooperates and defects), but they may also just get 1 on average (if they both defect). Although in a repeated situation, everybody is best-off in a cooperative society, where everyone always cooperates, this is hard to achieve. Theoretical or mathematical analysis of this game does not yield cooperating behavior, whereas we know from our own daily experience that cooperation does occur often indeed and proves to be advantageous.

Technique

In an adaptive social simulation we implement a society of agents in the form of a software simulation. We can do this by a so-called evolutionary algorithm (see Inset and Section 6.2.15, Evolutionary methods). The behavior of an agent is encoded here as a strategy addressing correct behavior, depending on the previous moves of the other agent. This strategy is used, evaluated, and changed in the adaptive social simulation together with other strategies, thus simulating the learning of the society of agents through time.

Results

When executing such a simulation, periods of cooperation and defection appear to alternate; in such periods most of the agents have similar (cooperative or defective) behaviors. Periods of cooperation emerge when more and more agents benefit from a higher pay-off when mutually cooperating. After a short while, however, 'exploiters' start to arise, who exploit the cooperative society by treason, i.e. defecting. When too many exploiters exist, the cooperative society breaks down and a period of defection starts.

Important extensions

In an adaptive social simulation, social conditions and incentives for cooperation can be tested and added. An important example is the usage of tags of agents: visible marks that essentially are meaningless, but after a while may get a meaning in the society of agents. For example, a tag may be comparable to clothes and fashion in real life, as in the following. Distinct social groups of people exist, each having a common behavior, interest, or profession. Such groups may want to distinguish from other groups in their appearance, e.g. by specific types of clothes or specific types of cars. Although meaningless items on their

own, such visible items can become indications of behavior and guide the interaction between people. In our case of the IPD game, a tag could get a meaning like: 'I am willing to cooperate'. A tag can thus give rise to a social or economic group with specific behavioral properties.
The usage of tags appears to substantially ease and improve the emergence of cooperative societies in the adaptive social simulations. Still, break-downs may occur as before, when the exploiters are appear as 'mimics': exploiters that mimic a cooperative impression by using a tag that has the meaning of intending cooperation. If too many mimics exist for a specific 'cooperative' tag, the tag will loose its meaning, and the cooperative group breaks down. So, although tags do not guarantee a fully stable cooperation, they do substantially increase the level of cooperation and the frequency of its occurrence.

Concluding

The influence of 'tags' on cooperation is hard to treat theoretically or mathematically. An adaptive social simulation does work out for investigation and derivation of properties, however [Axelrod, 1984; Alkemade, 2000]. In this way, important aspects of daily life can be incorporated into scientific models and business simulation systems. Important examples arise in commerce, marketing, economics, systems of software agents, and social sciences.

Inset: an adaptive social simulation technique

We describe a current computer learning technique that is suitable for adaptive social simulations: evolutionary algorithms [Mitchell, 1998]. Evolutionary algorithms (EAs) are strongly inspired by the genetic evolution theory in biology, as developed by Darwin. EAs typically work as follows, for solving a problem or simulating an agent system. First, a random 'population' (set) of possible candidate solutions for the problem are generated; the population can also consist of behavior strategies for the agent system considered. This population is subsequently changed and improved in a number of rounds ('generations') by means of evolutionary concepts. These concepts are 'survival of the fittest', selection, mutation, and recombination (like 'crossover'). This is repeated several times, for instance, for 1000 generations. The final typical result is a population with solutions or strategies that are as good as possible for the considered problem or system, and from which the (almost) best solutions can be selected. More in Section 6.2.15, Evolutionary methods.

How can emergence results be used?

From the emergent behavior of an agent system, behavior strategies for the individual agents can be developed. This means that by looking at the effects of various types of strategic behavior (e.g. from the business point of view), this behavior can be optimized according to the desired results. Also, the collective

behavior of many autonomous agents working together can be observed. This allows us to study, guide and control the development of an agent system. Especially, when software agents (computers) use adaptive or learning techniques to optimize their behavior, a question is whether these could give rise to unexpected outcomes. One could also think of the behavior of agents in future, dynamic markets: how will such markets behave? Can these be predicted or controlled? Obviously, it is desirable to know such possible behavior is for design and control purposes, as well as before actually making systems or agents operational.

Research activities and future developments

The focus on problems arising from the emerging ICT society is, of course, a very recent one. The work is an extension of pioneering work that was mainly done in the USA, at institutes such as the Santa Fe Institute [Holland, 1991; Holland, 1995] or the MIT Multimedia Lab [Maes, 1999]. At the former institute, a basis for considering economical systems as evolutionary systems was laid, while in the latter the role of software agents in commerce was explored. Currently, attention on these fields is growing in magnitude in the economics and computer science disciplines, and in the E-commerce and artificial intelligence disciplines respectively. An important instance is the field of Agent-based Computational Economics (ACE) [Tesfatsion, 2001] within economics and computer science. A recent example in this field is research carried out at the CWI in Amsterdam, addressing the development of adaptive social simulation systems for electronic markets.

To illustrate this, in order to model real-world agents in a market, realistic learning and interaction techniques in simulations are needed. So, various learning and adaptation techniques must be used and further developed to simulate the behavior of economic markets and agent systems. Also, the development and usage of other new techniques for adaptive social simulation is important, in order to address the different types of learning and social interactions, like for example reinforcement learning. In the future, the derivation of more formal theories of emergences from adaptive social simulation will also arise, as suggested in [Holland, 1995]. In this way, several results of emergent behavior in adaptive social simulations can be (re)used together. The types of research that are needed in these areas concern applicable research as well as fundamental research. This is because for many of the anticipated application areas, only a limited amount of fundamental research has been performed as yet.

For the actual application, cooperation between computer scientists and experts in the application fields is essential. In current research there is strong interest in multi-agent systems, as they can occur in economics, electronic commerce, business sciences, logistics, social systems, and ICT.

The final aim is to effectively and efficiently obtain emergent behavior proper-

ties for complicated agent systems, either in economics, social sciences or ICT. We can thus use this for, for example: derivation of market behavior, development of marketing or other strategies, design of market mechanisms, observation of social behavior in societies, or design and control of agent systems. The development of more and more powerful computers could thus give us powerful tools for economists, marketeers, designers of agent and ICT systems, or sociologists, for testing ideas and thoughts, and for developing proper solutions.

References

- Alkemade, F., D.D.B. van Bragt, H. La Poutré. (2000). Stabilization of Tag-Mediated Interaction by Sexual Reproduction in an Evolutionary Agent System. Proceedings of the First International Workshop on Computational Intelligence in Economics and Finance CIEF'2000. Vol. 2. Atlantic City, USA, pp945-949
- Axelrod, R. (1984). The Evolution of Cooperation. Basic Books, New York
- Bragt, D.D.B. van, C.H.M. van Kemenade, H. La Poutré. (2001). The Influence of Evolutionary Selection Schemes on the Iterated Prisoner's Dilemma. Computational Economics
- Dawid, H. (1996). Adaptive Learning by Genetic Algorithms: Analytical Results and Applications to Economic Models. Springer Lecture Notes in Economics and Mathematical Systems **441**
- Gerding, E., D.D.B. van Bragt, H. La Poutré. (2000). Multi-Issue Negotiation Processes by Evolutionary Simulation: Validation and Social Extensions. Proceedings of the Workshop on Complex Behavior in Economics: Modeling, Computing, and Mastering Complexity. Aix en Provence, France
- Holland, J.H., J.H. Miller. (1991). Artificial Adaptive Agents in Economic Theory. American Economic Review. Proceedings of the 103rd Annual Meeting of the American Economic Association. pp365-370
- Holland, J.H. (1995). Hidden Order: How Adaptation Builds Complexity. Addison Wesley, New York
- Kauffman, S.A. (1993). The Origins of Order: Self-Organisation and Selection in Evolution. Oxford University Press
- Maes, P., R.H. Guttman, A.G. Moukas. (1999). Agents that Buy and Sell. Communications of the ACM **42**:81-91
- Mitchell, M. (1998). An Introduction to Genetic Algorithms. MIT Press, Cambridge, Massachusetts
- Testfatsion, L. (2001). Structure, Behavior, and Market Power in an Evolutionary Labor Market with Adaptive Search. Journal of Economic Dynamics and Control **25**:419-457
- Vriend, N.J. (2000). An Illustration of the Essential Difference between Individual Learning and Social Learning and its Consequences for Computational Analyses. Journal of Economic Dynamics and Control **24**:119

– Weiss, G. (ed.). (1999). Multi-Agent Systems, A Modern Approach to Distributed Artificial Intelligence. MIT Press, Cambridge, Massachusetts

2.3.8 Data mining in environmental sciences

Monica Wachowicz[1]

Introduction

Environmental Sciences study the principles of reducing adverse effects of solid, liquid, and gaseous discharges to land, water, and air that can be responsible for degrading environmental resource values. Environmental databases are generated and frequently used for policy decisions, strategic planning, and research on global climate change, natural hazards, and land degradation. The environmental data being generated today originates from Earth Observation Systems, field measurements, model calculations and several other sources. The spatial dimension of these data is, in fact, the primary focus of analysis for studies of pollutant dispersal, forest fragmentation, and others. The temporal dimension is given by repeated observations that are critical to answering the most important environmental questions, those related to global driver indicators in sustainable development, climate change, land use management, and environmental processes (e.g. land-atmosphere interactions, biogeochemical processes, and hydrological processes). Therefore, environmental databases typically have spatial as well as temporal dimensions.

We are currently facing problems related to the greatly increased volume of environmental databases due to the improvements in data acquisition, validation, archiving, and distribution (e.g. instruments, sensors, computational resources, Internet). There is a clear need to respond to new data analysis challenges posed by the overwhelming volume and high resolution data sets generated today. Remotely sensed data from Earth Observation Systems alone is projected to yield 1 terabyte per day, far more than can be analyzed by conventional means. Further, the environmental data sets show a high variability in data formats, scale and content. They are also becoming more complex, partly as a result of the high dimensionality of these data.

Moreover, having the right environmental data set is insufficient to aid the formulation and monitoring policies required for improving the environment. Neither can data in itself be sufficient in responding to a variety of complex issues and their interrelations concerning the support for sustainable development and processes. Examples can be found in issues related to global concerns, regional disparities, and local implications. We need to go beyond the delivery of data to the delivery of information and knowledge derived from these data. Therefore, data mining methods and tools are of fundamental importance for environmental sciences. Without a systematic effort to generate data mining solutions, the environmental databases being created today will be greatly under-exploited, and our efforts to develop a data-information-knowledge-decision strategy will be in vain.

1 Dr M. Wachowitz, M.Wachowicz@Alterra.wag-ur.nl Wageningen UR, Centre for Geo-Information, Wageningen, The Netherlands, http://www.geo-informatie.nl

The definition used for data mining in this chapter will be the one raised during the last NASA workshop on the Issues in the Application of Data mining to Scientific Data (Behnke, 1999).
The definition states that data mining involves the "science, tools, environment, and facilities to scale up and or automate scientific analysis of large-scale data streams, consisting of:

- Exploration of anomalies in geophysical data, where the detection of an anomaly may initiate an 'alert' requiring further human-in-the-loop analysis (e.g. using statistical or other methods).
- Scaling up of current analysis techniques to detect known phenomena such that large-scale data product streams may be automatically analyzed.

And characterized by:

- Critical partnerships between physical scientists, computer scientists, and statisticians for the effective integration of analysis processes, scientific algorithms, statistical approaches, and enabling computer architectures." (Behnke, 1999).

Data mining techniques have been used in pursuit of one of the general tasks of clustering, classification, generalization and prediction. These tasks usually improve the quality and effectiveness of a decision-making process, mainly because they complement and can often replace other decision-assistance techniques, such as statistical analysis, data reporting and querying. This chapter proposes the data-information-knowledge-decision strategy based on new technology developments such as data mining techniques (see Figure 1). This strategy works on the assumption that policy, however devised, works better when policy makers and stakeholders are better informed. An information infrastructure and a knowledge base will support this strategy by using data mining techniques as a powerful and user-friendly tool for analysis given the overwhelming data volume of environmental databases. A multi-stakeholder interaction will allow non-specialist and interdisciplinary or cross domain users (stakeholders, policy makers, researchers) to share their knowledge on economical, technical and political feasibility as well as social acceptance.

Finally, the success of applying this strategy depends on the formation of coherent interdisciplinary teams involving users, systems developers and scientists. Teaming up in a wide range of technology development and interdisciplinary activities is crucial to its success, and must be closely coordinated with any sponsoring activities by governmental organizations and other funding agencies. The following Sections provide a proposal for a data-information-knowledge-decision strategy. This strategy would allow us to respond to increasingly complex, global and unexpected environmental issues. To achieve this, the methods used have to be able to cope with larger sets of criteria, parameters and quan-

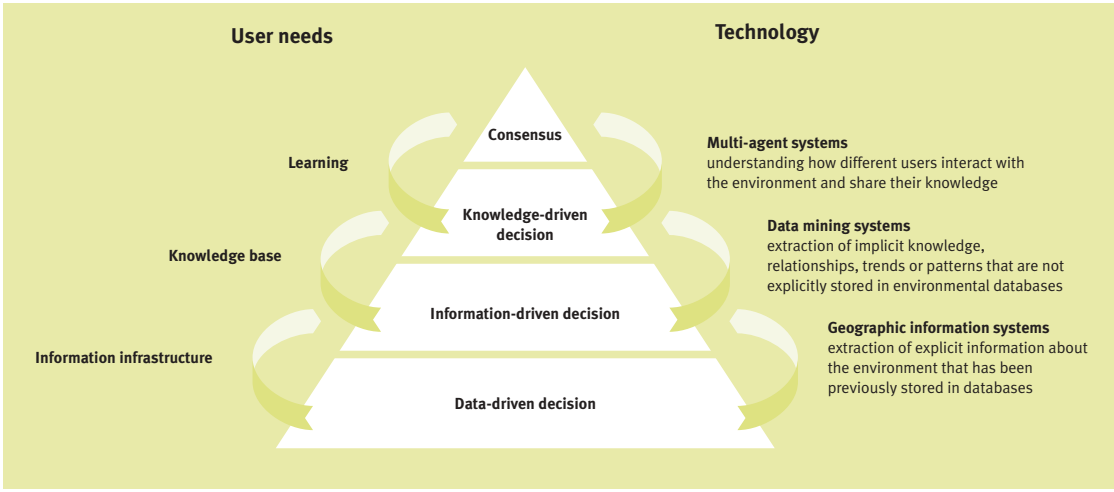


Figure 1
Overview of the data-information-knowledge-decision strategy.

tifiable indicators for sustainable development. Data mining techniques will provide the tools for supporting this strategy in order to discover association rules, prominent classes or clusters, spatio-temporal relationships, and others. A brief overview of the state of the art is given by describing what has been done to date. Finally, this Chapter concludes with what are considered the main research challenges in the near future.

Data: accessibility and interoperability

The first step of data mining involves discovering relevant data which can be mined for particular purposes. Environmental data sets are usually from different sources and they are collected for multi-purpose use, having different spatial and temporal scales, accuracy, and map thematic classes. Some types and characteristics of the available environmental data sets are described below (more details can be found in [Tateishi, 2000].

- *Digital topographic data sets* are often produced based on cartographic base map data, and used as sources for computing parameters such as production and or analysis of slope, aspect, hydrologic or process models at local or regional scales. The main scientific or technical applications for these data sets include environmental impact assessments, surface hydrologic flow and drainage basin delineation, atmospheric circulations at local, regional or global scales, and site navigation for aircrafts and missiles. The data sets are used by a variety of users such as government agencies (management), private citizens (recreational purposes), private companies (development) and educational institutions (research).
- *Digital elevation data sets* are mosaics of local or regional models of the digital elevation model of the Earth's surface. Digital topographic data, digital elevation data, and digital elevation model (DEM) are usually considered to be synonymous generic terms describing the attempts to describe elevations

and physiography. They are mainly used in hydrological models. Many countries have developed, and are currently producing and distributing gridded DEMs at a variety of scales. These DEMs are developed from hardcopy contour maps, using manual or scanner digitizing. Satellite altimetry such as SPOT satellite series and radar imagery is used for measurement of ocean heights, areas over ice caps and flat areas. Recently, laser altimetry is being used to produce better DEMs. Errors in elevation data may originate from preprocessing (occasional), from source materials or the conversion to digital form (systematic) or occur at random.

- *Oceanographic data sets* include a diversity of parameters such as bathymetry, coastal processes, marine geophysics, marine sediments, ocean acoustics, pressure, temperature, and water quality. Great importance was given to the development of standards in relation to data collection. However, data format standards have had less attention. Adoption of an agreed 'data dictionary' remains one of the most important tasks and challenges facing the data management community. We have a large amount of oceanographic data today that is frequently duplicated in different data structures and varying levels of data quality. The database used often depends on the need and the availability of data sets with the 'closest fit'.
- *Land cover data sets* play a vital role in environmental sciences, including studies of land-atmosphere interactions, biogeochemical processes, net primary productivity, and hydrological processes. The data sets are from coarse resolution data such as AVHR (Advanced Very High Resolution Radiometer) images, as well as intermediate resolution data from Landsat and SPOT images. The assessment of thematic accuracy of land cover data sets and derived products have generally received insufficient attention. Currently, there is a lack of consensus on optimal methods for the assessment of the thematic accuracy of products derived from remotely sensed data. The implementation of current methods can also be costly. A unified classification scheme for the land cover types has yet to be generally accepted.
- *Biodiversity data sets* include information on habitats, biotopes, ecosystems, species data, genetic data and biological reference collection. Data sets are vast, and it is extremely difficult to generalize their findings. However, the majority of the data are increasingly available in digital form, most of them concentrated in developed countries. Adoption of a consistent methodology for capturing the data still remains one of the inconsistency problems in the production of these data sets. The users include national institutions, regional groupings, international institutions, international funding agencies, bilateral development agencies, international environmental and conservation groups and scientific communities.

Table 1 (opposite page)
Overview of the National Clearinghouses for Geo-Information (17 March 2000) [Crompvoets, 2000].

Country name	Web page address	[1]	[2]	[3]	[4]	[5]	[6]
Belgium	http://www.vlm.be/OC/welcome1.htm (Flanders)	prototype	1998	150	n.f.	CEN TC/287	Dutch (English)
Denmark	http://www.daisi.dk/ http://www.geodata-info.dk/	developing	1997	180	104	CEN TC/287	Danish/English
Germany	http://www.ddgi.de/ http://www.atkis.de/	developing	1997	2,657	n.f.	CEN TC/287	German/English
Finland	http://www.nls.fi/ptk/infrastructure/index.html	initial	–	–	n.f.	–	Finnish
France	http://www.cnig.fr/	developing	1995	105	1,082	NF52000	French
Hungary	http://www.fomi.hu/hunagi/	initial	–	–	52	–	English/Hungarian
Ireland	http://www.tcd.ie/Geography/GIS/Geoid/	developing	1999	237	152	other	English
Italy	http://195.110.158.111/index.html	initial	–	–	492	–	Italian/English
Luxembourg	http://www.etat.lu/ACT/acceuil.html	initial	–	–	n.f.	–	French
The Netherlands	http://www.ncgi.nl/	developing	1995	1,533	1,070	CEN TC/287	Dutch
Austria	http://www.ageo.at/	initial	–	–	615	–	German
Poland	http://www.wloc.ids.pl/wodgik/sieci/gispol/edzia_g.html	initial	–	–	69	–	English/Polish
Portugal	http://www.cnig.pt/	developing/mature	1994	4,263	1,725	CEN TC/287	Portuguese (English)
Russia	http://www.fccland.ru/	initial	–	–	249	–	Russian/English
Slovenia	http://www.sigov.si:81/index-1.html	developing	1997	407	535	CEN TC/287	Slovenian
Spain	http://mercator.org/aesig/	initial	–	–	n.f.	–	Spanish
Czech Republic	http://labgis.natur.cuni.cz/cagi/	prototype	1998	120	n.f.	CEN TC/287	Czech
United Kingdom	http://www.ngdf.org.uk/	developing	1999	2,103	2,250	ISO TC/211	English
Iceland	http://www.hi.is/pub/gis	initial	–	–	269	–	English
Sweden	http://www.uli.se/	developing	1998	2,398	550	other	Swedish
Switzerland	http://www.sogi.ch/	initial	–	–	n.f.	–	French/German/English

[1] Phase description1: 4 different levels are distinguished: a. Initial (not built an actual 'Internet'-clearinghouse). b. Prototype (a built Internet-application', however not completely operational) c. Developing (clearinghouse with only access to meta-data files) d. Mature (clearinghouse with access to the 'real' data).

[2] Year, first implementation version.

[3] Number of data sets: contact 'webmaster' clearinghouse or counted.

[4] Number of visitors per month: contact 'webmaster' clearinghouse or read 'counter'. Number of references: use the http://www.wsabstract.com/linkcheck/index.htm 'Website Abstraction Link Popularity Checker'. Access to the following search engines Hot Bot, Alta Vista, en Info Seek. Language: (English) Site literature can be obtained in English. However, the content and search engines are not in English. (n.f. = not found, not applicable).

[5] Standard.

[6] Language.

Very few environmental data sets are available through the search portals of data suppliers or providers, and as a result, there is no central repository for environmental data that could be used to perform data mining. Most of these portals provide meta-data information, and there is still a need to develop a web or application server to access the actual data sets. User protocols and procedures are needed for providing different access rights, data formats, personalized interfaces for different types of users, task-group virtual folders, decision support rules, and user feedback information. Scalability methods will be required due to the large volumes of the data that need to be accessed. For example, the current situation of national geographic clearinghouses confirms that Internet access is only available for meta-data information (see Table 1).

Although there are differences in development stages, scale, country, standards, and types of phases, the clearinghouse development in Europe has not yet led to an intensive market on the Internet. The number of data sets available in a national clearinghouse varies from 4,263 data sets in Portugal to 150 data sets in Belgium. In contrast, there are approximately over 2,000.000 data sets available over the Internet in the USA. Moreover, the low numbers of visitors per month shows that providing only meta-data information has not encouraged free flow of data. Table 1 shows that only one national clearinghouse in Europe has over 2,000 visitors per month. In contrast, the average number of visitors per month is over 15,000 in the USA. One of the main reasons for this is due to the fact that American clearinghouses usually provide the actual data sets and operational applications (services).
In the future, the ultimate goal will be to provide mining interoperability of disparate data sets provided by an array of different data providers, sensors, and instruments. All data should be virtually on-line to allow a data mining system rapid access to the data. This may require the development of an open reference architecture consisting of object models for data and associated meta-data standards. It is expected that standardization in the field of environmental sciences will promote 'intelligent' interoperability, thus broadening a potential market for data mining applications.

Information: mining infrastructure and systems

Integrated policy making cannot exist without an information infrastructure that provides spatio-temporal references, remote access to environmental data, and easy ways to explore these data with interactive tools. The 'Bridging the Gap' Conference in June 1998 also pointed out the new needs and perspectives for environmental information, and its chairman stated that: "At present some of the systems for monitoring and gathering information about the environment in European countries are inefficient and wasteful. They generate excessive amounts of data on subjects which do not need it; and they fail to provide timely

and relevant information on other subjects where there is an urgent policy need for better focused information, and for consistent environmental assessment and reporting." [EEA 1999]. Environmental information is mainly needed to support policy initiatives in areas such as regional development, transport, environmental protection, agriculture and forestry.

The mining infrastructure concept has an immediate intuitive appeal, appearing to facilitate and promote the sharing of environmental information. However, it also brings several issues concerning a variety of behavioral or cultural, legal, economical, and organizational factors that can hinder the development of such an infrastructure. These factors pose conflicting demands on the organizations involved in the acquisition, development, custody, and dissemination of data. In addition, there are issues related to public access, intellectual property rights, data protection and security, liability and privacy. Naturally, economic factors are related to the complexity of financing and pricing strategies for achieving the dissemination of data.

One of the main consequences of this current situation is the difference between available information systems (e.g. GIS) and the expected data mining systems (e.g. GeoMiner, CONQUEST, AdaM). For example, the current focus of GIS is on providing a solution to given problems, and not on discovering knowledge per se. On the other hand, many methods implemented in data mining systems provide learning, interactive, exploratory, and visualization 'tools'. In fact, data mining tools can help us to extract useful patterns, objects, events, categories, and structures from databases. These can be used to construct a model of relationships between patterns and processes or events they represent. In environmental sciences, scientists are interested in using data mining to make a valuable contribution to improve our understanding of complex process characterizations. They would like to understand for example, how ocean, atmosphere, and land processes are coupled, how to deal with multiple time series from multiple sensors or instruments for a single purpose, and finally, what is the impact of human influences related to these processes. The forthcoming data mining systems will play an important role as the means of achieving these expectations.

Current data mining systems consist of tools that can perform a range of tasks. There are over 200 such tools currently available in the public domain[2], ranging from data preparation, classification, visualization, and web mining, to clustering and other forms of mining tasks. However, they have so far not had much impact on environmental sciences due to the spatial and temporal components of the environment data. The methods implemented in these tools are often not 'spatially aware' and where they are, they use very simple data models of spatio-temporal objects and relationships, for instance, snapshots of point object and Euclidean distances. Other complex spatio-temporal objects (e.g. moving

2 see http://www.kdnuggets.com/software

points, lines, polygons) and their respective relationships (e.g. direction, connectivity, non-Euclidean distances) also need to be integrated into a data mining system. This will require a full range of conceptual, logical, and physical database models of spatio-temporal objects.

One of the most common findings in literature is that time is just one more dimension to be added to the spatial dimension [Wachowicz, 2000]. This perspective is indeed the underlying rationale behind most of the implementations of spatial and temporal database models. However, the synergy of space and time requires '*spatio-temporal concepts*' that represent space-time dynamics (for example, patterns and process of change) and the '*human cognition of a knowledge domain*' (for example, distinction between observed spatio-temporal patterns and derived knowledge). Adding the time dimension to a spatial data model is inadequate for representing space and time in databases. This is mainly because such an addition will result in a database model that represents the time dimension in the same manner as the spatial dimension, and as a result, it may only capture time-referenced sequences (snapshots) of spatial data.

Knowledge: data mining application areas

The identification of application areas to which data miners could target their research is still an open question. The final report on 'Issues in the Application of Data mining to Scientific Data' (Behnke, 1999) provides the first step towards the creation of an overview of potential application areas. This overview describes several examples of current and potential application areas, which have been aggregated into three major categories: spatial mining, temporal mining, and spatio-temporal mining.

Spatial data mining (mining using the spatial dimension):

- Current application areas: land cover mapping; eddies and fronts detection; latitudinal variation in Rossby radius[3] detection in ocean data; event detection such as cyclones; fires, and meso-scale connective systems; and cloud identification.
- Future application areas: mining radar data for storms, SeaWiFS[4] data for primary production, global characterization of land cover, distribution of carbon in the terrestrial ecosystem and in the ocean.

Temporal data mining (mining using the temporal dimension):

- Current application areas: mining surface changes over time (e.g. earthquake rupture), identification of growing season anomalies.
- Future application areas: mining time series using different sources of data including GPS[5], InSAR[6], seismic data, topographic data, etc.

3 The effective length of the lateral distance between the region of disturbance generation and its outermost extent: LR = (gh)1/2f, where f is the Coriolis parameter; g, the acceleration due to gravity; and h, the height.

4 Sea-viewing Wide Field-of-view Sensor (SeaWiFS) provides information that can be used to investigate biological productivity in the ocean, marine optical properties, the human influence on the oceanic environment, etc.

5 Global Positioning System.

6 A software application designed to produce digital elevation models and height change maps through the use of repeat pass SAR interferometry (Synthetic Aperture Radar).

Spatio-temporal data mining (mining using both spatial and temporal dimensions):

- Current application areas: this is the most complex area in data mining. Among the few examples available are: disease correlation between SST/AVHRR[7] data, association rules in land cover changes.
- Future application areas: this domain represents the majority of data mining application areas for the future. They include exploratory pattern mining (e.g. is the climate changing?), cause relationship (e.g. determine if land clearing has perturbed runoff and flood frequency), indicator relationships (e.g. identify whether terrestrial change detection can be indicator of global change), effect relationships (e.g. find synoptic events having regional climate impact such as volcanic eruptions, flash floods), and prediction relationships (e.g. predict storm tracts, droughts, floods or fire potential).

Consensus : agents for a decision-making process

It is almost impossible to achieve a decision-making process among different stakeholders, decision makers, and scientists without developing a knowledge construction process. Decision-making is in fact a construction of shared and personal representations of knowledge about a problem domain, which converge through the interaction of interpersonal relations. Therefore, it is fundamental to point out that in the perspective of modeling and supporting a decision-making process we need to assess how knowledge representations are built or exchanged and evolved. This conception of decision making as a knowledge construction process of reality emphasizes the contingency of different types of knowledge and the relativity of information as a constant background. Sustainable development is currently an extremely sensitive issue, especially in the Netherlands, where old traditional settlements, high population density within some critical areas, and complex political influences and relations, have led to an environmental awareness regarding land use activities. Many studies and developments of models for environmental assessment, multi-criteria analysis, expert practices, consensus and negotiation are available in the literature, but only few address the issue of including a knowledge construction process. For example, [Ferrand, 1996] emphasizes how modeling a decision-making process is about how the world is perceived and communicated, rather than how it really is. This perspective supports the data-information-knowledge-decision strategy discussed in this chapter. In this strategy, decision-making consists of the evolution of beliefs, knowledge, and preferences, and how these components will determine a person's behavior.

From this perspective, data mining systems play an important role as a tool to construct different types of knowledge from large environmental databases. But this perspective will also rely on the integration of data mining techniques

7 SST-Sea Surface Temperature data and AVHRR-Advanced Very High Resolution Radiometer data.

with other new technology developments such as multi-agents systems (see also Paragraph 2.3.7, Emergence).
"A multi-agent system is a set of agents interacting in a common environment, where an agent is a computing element executing its design goals in this environment and able to modify both its environment (communication, decision, action) and itself (perception, reasoning, learning)." [Ferrand 1996; Weiss, 1999]. Multi-agent systems can differ in terms of the agents themselves, the interactions among agents, the agent architecture used in the system, and the computer environments in which the agents act. For example, information agents can access multiple, potentially heterogeneous and geographically distributed environmental information sources. They can cope with the increasing complexity of modern information environments, ranging from relatively simple in-house information systems, through large-scale multi-databases systems. One of the main tasks of these agents is an active search for relevant information in non-local domains on behalf of their users or other agents. This includes retrieving, analyzing, manipulating, and integrating information available from different information sources.
In contrast, an interface agent is typically a software agent that supports its user(s) in fulfilling certain tasks. For instance, an interface agent may hide the complexity of a difficult task, train and teach a user, and perform subtasks on a user's behalf. The terms software assistant and personal assistant are often used in this sense. Interface agents also play an important role in computer supported co-operative network.
Decision making is an inherently user-specific process in which every user requires a specific set of services and analytical tools. The data-information-knowledge-decision strategy is the first step towards the development of a decision making process in which different users identify their needs, define the planning situation, learn about their plans, and make decisions based on these plans.

State of the art

Data mining tasks

Typical tasks for data mining are clustering, classification, generalization and prediction.

Clustering

Clustering (see 6.2.5 Clustering) is probably the most widespread *mining task* being developed in the field of environmental sciences. Clustering (also called segmentation) is the task of partitioning a database into several sets (clusters) in such a way that all members of a set are similar according to structural similarity criteria or rules. Some examples of methods for clustering large data sets

using similarity rules are CLARANS - Clustering Applications based upon Randomised Search [Ng, 1994] and BIRCH (Balanced Iterative Reducing and Clustering [Zhang, 1996]. Moreover, SD-CLARANS (spatial dominant algorithm) and DSD-CLARANS (non-spatial dominant algorithm) are two extensions of CLARANS [Ester, 1996].

Classification

The *classification mining task* examines the features and places them into classes (categories) according to meaningful partition criteria, model, or rule See 6.2.7 Classification). Most of the classification algorithms are based on the induction perspective using symbolic and statistical methods. Symbolic methods address the issue of producing sets of statements about local dependencies among features in a rule form (See [Chen, 1996] for an overview of classification algorithms based on symbolic methods). On the other hand, statistical methods focus on exploiting statistical discriminators (probability distributions, hypothesis generation, model estimation and scoring) for extracting categories from a data set using supervised/unsupervised learning, cluster analysis, and related methods [Hosking, 1997].

Generalization

The *generalization mining task* consists of finding a concise and condensed description for a database. The goal is to provide users with multiple perspectives on data, thus allowing them to detect features that may exist only at a particular conceptual level (i.e. levels of abstraction). This mining task has much in common with cartographic generalization - particularly as cartographic generalization has already been used for building multi-resolution databases [Frank, 1994]. Some examples of data mining systems that perform a generalization task are DBMiner [Han, 1996] and GeoMiner (an extension of DBMiner for applications to geo-referenced data [Han, 1997]. GeoMiner allows users to examine different levels of detail of a generalized hierarchy. Summarized tables, charts, and maps are also employed to create snapshots of the generalized hierarchies. Generalization can also be seen as modeling or rule extraction.

Prediction

Prediction is the *data mining task* that has attracted the highest level of interest due to the enormous benefits expected from the outcomes of predictive modeling. In particular, the immense surge of interest in abstracting and predicting storm tracts, potential conditions that will result in potential droughts, floods or fire. At a more complex level, scientists are interested in applying prediction data mining techniques to evaluating and predicting global sustainability.
For a view on data mining tasks from a business perspective , see Chapter 3.1, Introduction and for a technical look see Chapter 6.2, Techniques.

Interactivity through visualization

However, in most of the data mining systems, data is still considered static and the kind of knowledge to be mined is defined in terms of the mining task to be performed, such as clustering, classification, prediction, or generalization. The human computer interaction resembles that of traditional database manipulation, re-running of an algorithm, fine-tuning through a series of queries, or re-selecting a target data set. A poor or erroneous choice of data input, mining task or algorithm will be only perceived after the results are obtained at the end of the process.

Historically, especially in statistics, the term data mining has often been referred to sloppy exploratory data analysis with no a priori hypothesis to verify [Glymour, 1997; Fayyad, 1996]. In order to avoid that, data mining should *always* be considered as a human-centered process, in which users can dynamically interact with the system and take their analysis decisions at any stage of the knowledge construction process. For a long time we have made a distinction between methods that are essentially concerned with visualizing data, data mining, and those which rely on the specification of a database model and its implementation. By having this distinction between methods for visualizing, mining, and modeling, we have defined a clear-cut approach that is not useful. We are always interested in the accurate description of data relating to a process operating in space and time, the exploration of patterns and relationships in such data, and the search for explanations of such patterns and relationships.

Therefore, several distinct types of visual data mining and exploratory analysis techniques are needed. From the perspective of data mining, [Hinneburg, 1999] describes four: geometric projection, iconographic, pixel-based and hierarchical. When considering those techniques used by geographers and statisticians involved in exploratory analysis chart-based and map-based techniques can also be added. A brief description of some of the more common families of techniques follows; greater detail is given in [Gahegan, 2001; Kraak, 1999]. Notice that there is no consistency to the way these various groups are defined; some are named after their data representation methods (e.g. map-based and hierarchical techniques) and others by methods applied to this representation (e.g. projection techniques).

Map-based techniques

Map-based techniques allow the mapped data and its visual appearance to be changed interactively [Dykes, 1997]. Map legends are often used as the basis for interaction, as shown by [Peterson, 1999] and [Andrienko, 1999], permitting the user to change the appearance of the objects mapped and thereby define and possibly explain clusters. Chart-based techniques plot the data on a chart or graph, common examples being scatter plots and parallel co-ordinate plots.

Scatter plots use a simple 2D or 3D graph with dots or spheres to mark the position of individual data items [Cleveland, 1988]. Parallel co-ordinate plots employ a (usually larger) number of parallel axes through which a trace of each data item can be made [Inselberg, 1985; Inselberg, 1997]. These techniques are often accompanied by linking and brushing methods, allowing selected data points to be viewed in different ways or within different axes (e.g. [Buja, 1996; MacEachren, 1999; Edsall, 1999].

Projection techniques

Projection techniques use statistical transformations such as principal component analysis and multidimensional scaling to project structure or trends from the data [Asimov, 1985; Haslett, 1996; Cook, 1995]. They are also often based around graphs, particularly scatter plots, so most examples could be seen as building on the chart techniques defined above.

Pixel techniques

These techniques map data values to individual pixels that are ordered on the screen so that data streams of similar values produce visible clusters in 2D [Keim, 1996a; Jerding, 1998]. The screen can be divided into separate windows, if several attributes are to be visualized [Keim, 1994; Keim, 1996b] Such techniques may present a useful overview of a very large data set.

Iconographic techniques

Iconographic techniques use complex symbols, such as stick figures [Pickett, 1988] or faces [Chernoff, 1973; Dorling, 1994] to encode many data dimensions simultaneously. The aim of iconographic displays is to promote perception of the 'whole', while still allowing some differentiation of individual variables.

Hierarchical and network techniques

As the name suggests, these techniques organize data strictly, according to a specific data structure, such as a tree [Robertson, 1991] or network [Huffaker, 1999; GeoBoy®, http://www.ndg.com.au/], with progressive levels refining the display into subspaces.

For visualization techniques, see also Section 6.2.20 and 3d information visualization[8] on the CD-rom.

Research challenges

A visual approach to data mining is needed for the analysis of large environmental databases (from well-structured vector and raster models to unstructured models such as georeferenced multimedia data) with multi-agents tools and visualization functionality. This analysis can take many forms, the extremes being a monolithic, single system and a component-based, loose federation

8 CD-rom: ..\papers\3dvisualisation_Young.htm

[Rhyne, 2000; Slocum, 1994]. A visual data mining infrastructure, implemented using a Java component architecture is given by [Hao, 1999]. [Gahegan, 2000] shows how such architecture can provide useful exploration of the data, leading to improvements in category construction.

Visual approaches to data mining will put emphasis on the user rather than on the system. For example, pixel-based and projection methods tend to use pre-defined transformations applied to the data, and these also explicitly define a hypothesis by which any observed pattern might be explained. If the transformations are entirely pre-defined, in other words they cannot learn from or be changed by the specifics of the data under consideration, then the system is operating by deduction. The user may not be, however. What a user notices in a visual display and how he chooses to interpret it is not defined (although it may be severely biased). Alternative visualization methods, such as the linked views in scatter plots and other dynamic exploratory techniques, are needed to offer a less deterministic structure, since the data representation is controlled less by the system and more by the user (and possibly the data itself). This will provide greater flexibility, and less external bias to the perception and knowledge construction process.

Other system issues that may affect data mining relate to the richness of models and concepts in the underlying database(s), specifically, their organization and the facilities they can provide for representing knowledge once it has been discovered (e.g. [Sheth, 1997; Drew, 1998; Sowa, 1999]. The most important missing functionality in current databases is 'data modeling management', allowing users to update the model as the process of knowledge construction evolves and to monitor these changes over time as objects, categories and relationships are uncovered.

For example, when a decision tree is used to inductively learn a land cover category or predict a stream outflow, it is also generating new knowledge in the form of rules. This knowledge also belongs to the database, and as a result, the diagram must be updated to accommodate it. Technically, it is an enormous challenge to determine how to incorporate this functionality into a database.

Data mining systems will also require the development of mining infrastructures that will support data integration to facilitate the construction of the set of data to be mined using data relevant to the objectives of the users (i.e. decision makers, stakeholders, and researchers). This infrastructure should contain the necessary tools for finding and obtaining the data, dealing with multiple formats and spatio-temporal scales, making the mining results accessible, and dealing effectively with the computational intensity inherent in data mining. For certain environmental mining results, the infrastructure needs to support the rapid systematic dissemination of the mining results and warnings (e.g. location of zones with high risk of flooding).

Although data mining methods have mainly been developed for non-spatial data sets, some approaches have specifically concentrated on spatial, temporal or spatio-temporal data sets (e.g. [Koperski, 1996; Wachowicz, 2001]) and a useful on-line bibliography is provided by [Roddick, 1999]. This key area of concern relates to the nature of environmental data, and the intrinsic notion that location matters, because space forms the framework that both defines patterns and determines their significance. For many tasks, effectiveness will depend on the simultaneous presentation of attribute values and their spatio-temporal dimensions, forming some minimal context by which data is given meaning. This fact should affect the design of or selection of effective data mining methods; they must be able to compute at least four variables (*x*, *y*, *t* and *attribute*) for environmental data sets, preferably in an integrated manner. A deeper and related question involves the general problem of using only attributes to identify objects of interest. Humans use relationships and other expertise that extends beyond simple property values, when identifying objects or defining categories (e.g. [Rosch, 1975]). Most data mining methods do not have very sophisticated mechanisms to replicate this expertise, nor indeed is it straightforward to gather them from a human expert or represent them in a machine. Decision making is a knowledge construction process that reflects not the end result a user would want, or what the system should produce, but rather the operations a user will perform with the tools of a data mining system. This is an important distinction, because it places the user first in the knowledge construction process, but it also highlights the semantic 'gap' between desired outcomes and the available tools. Existing data mining systems focus on identifying what kind of knowledge the system needs to discover successfully. However, it is very important to take the design initiative of enabling the users' expertise within a knowledge construction process, rather than attempting to supplant it. The challenge here is to demonstrate empirically that a combination of human and machine 'intelligence' really does improve knowledge construction capabilities and to build environments that take advantage of the best that each has to offer. For example, the *a priori* knowledge of a domain expert is difficult to pass on to a data mining algorithm. Conversely, the results of data mining are biased by the search criteria used, but these too might be difficult to communicate to the expert.

Mining environmental data often requires the space dimension to be explicitly presented. The need for spatial information may also have consequences for data indexing and retrieval. The huge volumes of environmental data now available and the highly-multivariate nature of some of the inherent trends and patterns pose additional challenges, when designing useful data mining techniques. It is not clear how space should be represented, for example, by name of a city or its geographical co-ordinates. Similar options exist for time also, e.g.

as a linear scale (days of a year) or a cyclic scale (hours of the day, seasons of a year). Mappings between these different representations will be required in many cases.

Finally, the representation of spatio-temporal knowledge is problematic. At present, there are no universal languages for spatio-temporal representation, data is represented and manipulated by the ad hoc models developed by GIS and database vendors. Specifically, rich conceptual structures (e.g. [Rosch, 1975]) are all but absent from current, computationally based database models. The only structure regularly present is that of the category, although some more recent systems based on the object-oriented data model may also contain generalization hierarchies and component relations, for example 'is a part of' [Wachowicz, 1999]. Perhaps the first task is to specify a list of spatio-temporal-oriented concepts we wish to be able to identify or 'mine', and a means of representing them in a current GIS or database schema. Following from this, we must define computational and visualization methods to detect, observe and communicate them.

Acknowledgments

The author would like to thank Anne M. Schmidt and Arnold Bregt for their comments and suggestions on the earlier version of this chapter.

References

- Andrienko, G.L., N.V. Andrienko. (1999). Interactive Maps for Visual Data Exploration. International Journal of Geographic Information Science **13** (4):355-374
- Asimov, D. (1985). The Grand Tour: a Tool for Viewing Multidimensional Data. SIAM Journal of Science and Statistical Computing **6**:128-143
- Behnke, J., E. Dobbinson, S. Graves, T. Hinke, D. Nichols, P. Stolorz. (1999). NASA Workshop on Issues in the Application of Data mining to Scientific Data. Final Report. Goddard Space Flight Center, USA
- Brachman, R.J., T. Anand. (1996). The Process of Knowledge Discovery in Databases. In: U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy. (eds.). Advances in Knowledge Discovery and Data mining. AAAI/MIT Press, Cambridge, MA. pp37-57
- Buja, A., D. Cook, D. Swayne. (1996). Interactive High-Dimensional Data Visualization. Journal of Computational and Graphical Statistics **5** (1):78-99
- Chen, M., J. Han, P.S. Yu. (1996). Data mining: An Overview from a Database Perspective. IEEE Transactions on Knowledge and Data Engineering **8**:866-883
- Chernoff, H. (1973). The Use of Faces to Represent Points in K-Dimensional Space Graphically. Journal of the American Statistical Association **68**:361-36

- Cleveland, W.S., M.E. McGill. (1988). Dynamic Graphics for Statistics. Wadsworth & Brookes/Cole, Belmont, California, USA
- Cook, D., A. Buja, J. Cabrera, C. Hurley. (1995). Grand Tour and Projection Pursuit. Computational and Graphical Statistics **4** (3):155-172
- Crompvoets, J.W.H.C. (2000). Personal Communication
- Dorling, D. (1994). Cartograms for Human Geography. In: H.M. Hearnshaw, D.J. Unwin. (eds.). Visualization in Geographical Information Systems. Wiley, Chichester, England. pp85-102
- Drew, P., J. Ying. (1998). Meta-data Management for Geographic Information Discovery and Exchange. In: A. Sheth, W. Klas. (eds.). Multimedia Data Management: Using Metadata to Integrate and Apply Digital Media. McGraw-Hill. pp89-121
- Dykes, J.A. (1997). Exploring Spatial Data Representation with Dynamic Graphics. Computers & Geosciences **23** (4):345-370
- EEA. (1999). Environment in the European Union at the Turn of the Century. Report European Environment Agency, Copenhagen. Office for Official Publications of the European Communities, Luxemburg
- Edsall, R.M. (1999). The Dynamic Parallel Coordinate Plot: Visualizing Multivariate Geographic Data. Proceedings 19th International Cartographic Association Conference, Ottawa. http://www.geog.psu.edu/~edsall/JSM99/paper.htm
- Ester, M., H.-P. Kriegel, J. Sander, X. Xu. (1996). A Density Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proceedings 2nd International Conference on Knowledge Discovery and Data mining (KDD-96). pp226-231
- Fayyad, U., G. Piatetsky-Shapiro, P. Smyth. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of the ACM **39** (11):27-34
- Ferrand, N. (1996). Modelling and Supporting Multi-Actor Spatial Planning Using Multi-Agents Systems. Proceedings Conference on GIS and Environmental Modelling. NCGIA
- Frank, A.U., S. Timpf. (1994). Multiple Representations for Cartographic Objects in a Multi-Scale Tree - an Intelligent Graphical Zoom. Computer Graphics: Special Issue on Modelling and Visualisation of Spatial Data in GIS **18** (6):823-829
- Gahegan, M., M. Takatsuka, M. Wheeler, F. Hardisty. (2000). GeoVISTA Studio: A Geocomputational Workbench. Proceedings 4th Annual Conference on GeoComputation, UK. http://www.ashville.demon.co.uk/gc2000/
- Gahegan, M., M. Wachowicz, M. Harrover, T.M. Rhyne. (2001). The Integration of Geographic Visualization with Knowledge Discovery in Databases and Geocomputation. Cartography and Geographic Information Science **28**

(1):29-44
- Glymour, C., D. Madigan, D. Pregibon, P. Smyth. (1997). Statistical Themes and Lesson for Data mining. Data mining and Knowledge Discovery **1**:11-28
- Han, J., Y. Fu, W. Wang, J. Chiang, W. Gong, K.D. Koperski, Y. Lu, A. Rajan, N. Stefanovic, B. Xia, O.R. Zaiane. (1996). DBMiner: A System for Mining Knowledge in Large Relational Databases. Proceedings of International Conference on Mining and Knowledge Discovery (KDD96), Oregon, USA
- Han, J., K. Koperski, N. Stefanovic. (1997). GeoMiner: A System Prototype for Spatial Mining. Proceedings ACM-SIGMOD 1997, Arizona
- Hao, M., U. Dayal, M. Hsu, J. Baker, R. D'Eletto. (1999). A Java-Based Visual Mining Infrastructure and Applications. Proceedings InfoVis'99. October 24-29. San Francisco, CA. pp124-127
- Haslett, J., R. Bradley, P. Craig, A. Unwin, G. Wills. (1991). Dynamic Graphics for Exploring Spatial Data with Application to Locating Global and Local Anomalies. The American Statistician **45** (3):234-242
- Hinneburg, A., D. Keim, M. Wawryniuk. (1999). HD-Eye: Visual Mining of High Dimensional Data. IEEE Computer Graphics and Applications. pp22-31
- Hosking, J.R.M., E.P.D. Pednault, M. Sudan. (1997). A Statistical Perspective on Data mining. Future Generation Computer Systems **13**:117-134
- Huffaker, B., E. Nemeth, K. Claffy. (1999). Tools to Visualize the Internet Multicast Backbone. Inet'99 Proceedings. San Jose, CA. pp22-25. http://www.isoc.org/inet99/proceedings/4e/4e_3.htm
- Inselberg, A. (1985). The Plane with Parallel Coordinates. The Visual Computer **1**:69-97
- Inselberg, A. (1997). Multidimensional Detective. Proceedings IEEE Conference on Visualization (Visualization '97), IEEE Computer Society, Los Alamitos, CA. pp100-107
- Jerding, D.F., J.T. Stasko. (1998). The Information Mural: a Technique for Displaying and Navigating Large Information Spaces. IEEE Transactions on Visualization and Computer Graphics **4** (3):257-271
- Keim, D., H.-P. Kriegel. (1994). VisDB: Database Exploration Using Multidimensional Visualization. Computer Graphics and Applications. pp44-49
- Keim, D.A. (1996a). Pixel-Oriented Database Visualizations. Sigmod Record, Special Issue on Information Visualization
- Keim, D.A., H.-P. Kriegel. (1996b). Visualization Techniques for Mining Large Databases: a Comparison. IEEE Transactions on Knowledge and Data Engineering (Special Issue on Data mining)
- Koperski, K., J. Adhikary, J.J. Han. (1996). Knowledge Discovery in Spatial Databases: Progress and Challenges. Proceedings ACM SIGMOD Workshop on Research Issues on Data mining and Knowledge Discovery. Montreal, Canada. pp55-70

- Kraak, M.-J., A.M. MacEachren. (eds.). (1999). International Journal of Geographic Information Science: Special Issue on Exploratory Cartographic Visualization **13** (4)
- MacEachren, A.M., M. Wachowicz, R. Edsall, D. Haug, R. Masters. (1999). Constructing Knowledge from Multivariate Spatio-Temporal Data: Integrating Geographical Visualization with Knowledge Discovery in Database Methods. International Journal of Geographic Information Science **13** (4):311-334
- Ng, R., J. Han. (1994). Efficient and Effective Clustering Method for Spatial Data mining. Proceedings International Conference on VLDB. pp144-155
- Peterson, M.P. (1999). Active Legends for Interactive Cartographic Animation. International Journal of Geographic Information Science **13** (4):375-384
- Pickett, R.M., G.G. Grinstein. (1988). Iconographic Displays for Visualizing Multidimensional Data. Proceedings IEEE Conference on Systems, Man and Cybernetics. IEEE Press, Piscataway, USA. pp514-519
- Rhyne, T.-M. (2000). Scientific Visualization in the Next Millennium. IEEE Computer Graphics and Applications **20** (1):20-21
- Robertson, G.G. (1991). Cone Trees: Animated 3D visualization of Hierarchical Information. Proceedings ACM CHI'91. ACM Press, New Orleans. pp189-194
- Roddick, J.F., M. Spiliopoulou. (1999). A Bibliography of Temporal, Spatial and Spatio-Temporal Data Mining Research. SIGKDD Explorations **1** (1) (in press). http://www.cis.unisa.edu.au/~cisjfr/STDMPapers/
- Rosch, E. (1975). Cognitive Representations of Semantic Concepts. Journal of Experimental Psychology **104** (3):192-233
- Sheth, A. (1997). Data Semantics: What, Where and How? In: R. Meersman, L. Mark. (eds.). Database Application Semantics. Chapman and Hall. pp601-610
- Slocum, T.A., S. Egbert, C. Weber, I. Bishop, J. Dungan, M. Armstrong, A. Ruggles, D. Demetrius-Kleanthis, T. Rhyne, L. Knapp, J. Carron, D. Okazaki. (1994). Visualization Software Tools. In: A.M. MacEachren, D.R.F. Taylor. Visualization in Modern Cartography. (eds.). Pergamon, London. pp91-122
- Sowa, J.F. (1999). Knowledge Representation: Logical, Philosophical, and Computational Foundations. Brooks/Cole, Pacific Grove, CA, USA
- Tateishi, R., D. Hastings. (2000). Global Environmental Databases: Present Situation; Future Directions. ISPRS, Working Group IV/6
- Wachowicz, M. (1999). Object-Oriented Design for Temporal GIS. Taylor & Francis, London
- Wachowicz, M. (2000). How Can Knowledge Discovery Methods Uncover Spatio-Temporal Patterns in Environmental Data? In: B.V. Dasarath. (ed.). Data mining and Knowledge Discovery: Theory, Tools, and Technology II. Proceedings of SPIE **4057**:221- 229
- Wachowicz, M. (2000). The Role of Geographic Visualisation and Knowledge

Discovery in Spatio-Temporal Data Modelling. In: H. Heres. (ed.). Time in GIS: Issues in Spatio-Temporal Modelling. Publications in Geodesy **47**:13-26

– Wachowicz, M. (2001). GeoInsight: an Approach for Developing a Knowledge Construction Process Based on the Integration of GVis and KDD Methods. To appear in: H.J. Miller, J. Han. (eds.). Geographic Knowledge Discovery and Spatial Data mining. Taylor & Francis, London
– Weiss, G. (1999). Multi-Agents Systems. MIT Press, Cambridge, USA
– Zhang, T., R. Ramakrishman, M. Linvy. (1996). BIRCH: an Efficient Data Clustering Method for Very Large Databases. Proceedings ACM-SIGMOD 1996, Canada

2.3.9 Ecological informatics in river management

Peter Goethals[1]

Abstract

In computer sciences several tools with interesting applications in ecosystem management were developed during the last decennia. Due to the pressing need for information stream optimization tools in ecological management, 'ecological informatics' emerged as a quickly growing scientific cluster during recent years. Database management, development of predictive models and knowledge visualization are probably the main exponents of ecological informatics. All three components are discussed in this review with a major focus on the development and application of river ecosystem models in an educational perspective. Models that offer a prediction of faunal responses to changes in environmental features (e.g. changes in discharge regime, dissolved oxygen level, etc.) can be of considerable value for river management. The development process of models based on artificial neural networks to predict biological river communities is presented from data collection to model validation. Finally, two practical applications of models in river management are provided.

Introduction

Ecological informatics is a recent hybridization between ecological and computer sciences. Contemporary ecological management deals with large information streams of data on land use, biological communities, structural and morphological characteristics of rivers and landscapes, physical and chemical composition of ecosystem components and climate, to name a few. Information handling tools from computer sciences are very useful as a support to decision making. Examples are database development and maintenance, data mining, development of predictive models, data and information visualization. In this chapter a brief overview of ecological informatics tools are presented with regard to river basin management.

Ecological informatics for river conservation and restoration

Database development and management

Data and information are the basic products of scientific research. In ecological research, where field experiments and data collections are very expensive and time-consuming, data represent a valuable and often irreplaceable resource for scientific research and nature management.

Data management can be viewed as a process that begins with the conception and design of the research project, continues through data capture and analy-

1 P. Goethals, MSc., peter.goethals@rug.ac.be, Laboratory of Environmental Toxicology and Aquatic Ecology, Ghent University, Gent, Belgium.

sis, and culminates with publication, data archiving and data sharing with a broader public [Michener, 2000].

The design of an effective data management system depends on considerable forethought and planning to meet and balance several fundamental requirements or objectives. The primary goal of a data management system is to provide data of a requested quality (data reliability, number of missing values, etc.) within a reasonable budget. A second system requirement is facilitating access to data by investigators. An important related issue is providing short-term and long-term security for data through data archiving. Archival storage involves various activities that are designed to protect the data against information fuzzyfication and loss. A data management system may therefore have the following components or activities [Michener, 2000]:

- an inventory of existing data and resources will have to be compiled and priorities for implementation be set;
- data will have to be designed and organized by establishing a logical structure within and among data sets that will facilitate their storage, retrieval and manipulation;
- procedures will be required for data acquisition and quality insurance and quality control;
- data set documentation protocols, including the adoption or creation of meta-data content standards and procedures for recording meta-data, will need to be developed;
- procedures for data archival storage as well as maintenance of printed and electronic data will have to be developed;
- an administrative structure and procedures will have to be developed, so that responsibilities are clearly delineated.

Ecologists can avoid many potential difficulties in field sampling and subsequent data analyses, if sufficient thought is given to designing data sets prior to collecting data. Therefore preliminary sampling and or information research is often useful in setting up a definitive intensive monitoring campaign (see Case study 1). Some decisions about data design are necessary before data are collected in order to produce field and laboratory data sheets. The completed design can be transferred directly to data entry tools to aid in data collection, to facilitate analysis by statistical software and to support meta-data development and to structure the data set for archiving (see example in Case study 2).

Case study 1: Site description (the Zwalm river basin)

The Zwalm river basin is part of the Scheldt river basin [Carchon, 1997]. The Zwalm river basin has a total surface of 11,650 ha. The Zwalm river itself has a length of 22 km (Figure 1). The average water flow at Nederzwalm, very near the

Figure 1
The Zwalm river basin in Flanders (Belgium).

Scheldt is about one m^3s^{-1}. It has a very irregular regime, with low values in the summer (minima lower than 0,3 m^3s^{-1}) and relatively high values in rainy periods (maxima up to 4,7 m^3s^{-1}) [Lauryssen, 1994]. The water quality in the Zwalm river basin improved substantially during the year 1999 due to investments in sewage and wastewater treatment plants during the last years [VMM, 2000]. None the less, most parts of the river are still polluted by untreated urban wastewater discharges and by diffuse pollution originating from agricultural activities.

Although Flanders is in general rather flat, the Zwalm river basin is characterized by several height differences, resulting in a very unique river ecosystem within the Flemish region [Soresma, 2000]. Consequently, soil erosion is the most important geo-morphological process resulting in a substantial transport of (contaminated) sediments in the river [AMINAL, 1999]. Structural and morphological disturbances are also numerous [Carchon, 1997]. Weirs for water quantity control obstruct fish migration and are one of the main ecological problems within the river basin. For this reason an in-depth study has been made on the construction of fish migration channels. This study, which also covered natural overflow systems, is intended to achieve an ecologically friendly water quantity management in the near future [Soresma, 2000]. Some upper parts of the watercourses in the Zwalm river basin are colonized by very rare species as the Bullhead (*Cottus gobio*), the Brook Lamprey (*Lampetra planeri*) and several vulnerable macroinvertebrates.

Case study 2: Data collection

Structural characteristics (meandering, substrate type, flow velocity, ...) and physical-chemical variables (dissolved oxygen, pH, ...) were used as inputs to

predict the presence or absence of macroinvertebrate taxa in the headwaters and brooks of the Zwalm river basin (see Table 1). Structural characteristics were visually monitored and classified [Dedecker, 2001]. Flow velocity was determined by timing the transport of a float over a distance of 10 m. Field measurements were made for temperature and dissolved oxygen (TW OXI 330/SET), pH (Jenway 071) and conductivity (WTW LF 90). Suspended solids were measured in the laboratory based on spectrophotometric measurements [Dedecker, 2001].

Table 1
Monitored variables in the Zwalm river basin [Dedecker, 2001].

Variables	Units
pH	
temperature	˚C
dissolved oxygen	mg/l
conductivity	µS/cm
suspended solids	mg/l
water level	cm
fraction of pebbles	%
shadow	%
water plants	2 classes: 0 = absent; 1 = present
width	cm
flow velocity	m/s
meandering	6 classes (1 = well-developed to 6 = absent)
hollow river beds	6 classes (1 = well-developed to 6 = absent)
pools/riffles	6 classes (1 = well-developed to 6 = absent)
artificial embankment structures	3 classes (0 = absent; 1 = moderate; 2 = intensive)

Macroinvertebrates were collected by means of a standard hand net [NBN, 1984] during a five minute kick sampling. The sampling was aimed at collecting the most representative diversity of the macroinvertebrates within the examined site [Pauw, 1983]. The absence or presence of macroinvertebrate taxa was respectively represented by 0 or 1 for use in the different models. In total, 60 sites were monitored in the Zwalm river basin.

Ecological modeling

Nowadays, a large set of modeling techniques is available to develop models for a broad range of applications. Artificial neural networks (see Section 6.2.8, Neural networks for data mining) [Lek, 1999], fuzzy logic (See Section 6.2.16, Fuzzy logic techniques) [Barros, 2000], evolutionary algorithms (See Section 6.2.15, Evolutionary methods) [Caldarelli, 1998], and cellular automata [Gronewold, 1998], for example, proved to be powerful tools for ecological modeling, especially when large datasets were involved [Goethals, 2001]. In Case

study 3, an example of an ecosystem model is presented. In Case study 4 the validation of these artificial neural network models is presented, reflecting the reliability of the predictions on new scenarios calculated by the models.

Case study 3: Development of predictive models for macroinvertebrate communities based on artificial neural networks

Artificial neural networks (ANN) is a technique from the field of artificial intelligence. In this example, back-propagation [Rumelhart, 1986] was used. With this type of ANN a set of training examples consisting of an input and an output vector with data from the Zwalm river ecosystem, is presented to the network. The back-propagation network determines its own parameters with specific 'training algorithms'. After training, the neural network is able to calculate an output vector for any new input vector, which allows the calculation of predictions for different scenarios in the described ecosystem. The setup of the applied neural network is given in Figure 2.

Artificial neural networks were applied to predict macroinvertebrate communities in the Zwalm river basin. Structural characteristics and physical-chemical variables mentioned in Case study 2 were used as inputs to predict the presence or absence of macroinvertebrate taxa and the Belgian Biotic Index [Pauw, 1993] in the headwaters and brooks of the Zwalm river basin.
All neural networks were implemented using the neural network extension of the software package MATLAB 5.3 for MS Windows™ [Demuth, 1998].

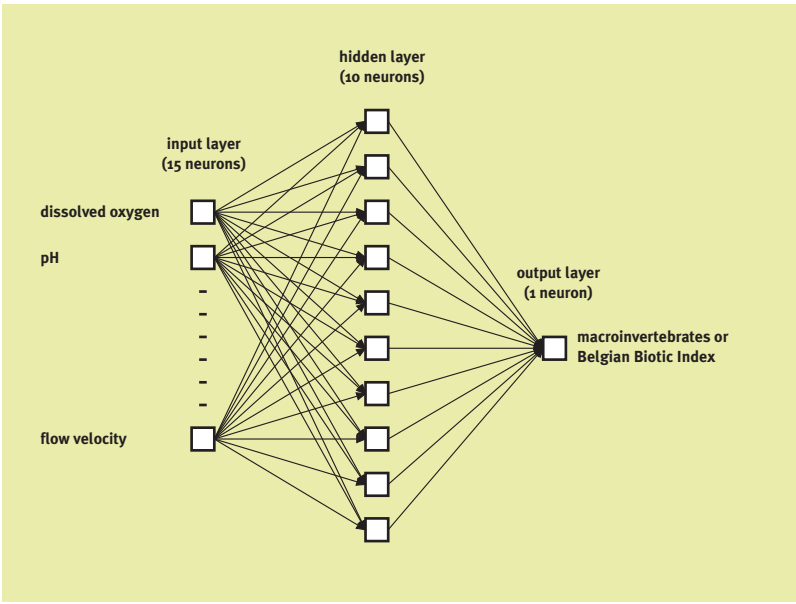


Figure 2
Scheme of the ANN model used for the prediction of macroinvertebrate taxa and the Belgian Biotic Index.

Case study 4: Artificial neural network models validation

To calculate reliability, the models were offered a dataset of 20 physical-chemical and structural measurement sets from the Zwalm river to predict the corresponding macroinvertebrate communities. These predicted macroinvertebrate communities were compared with the field data. In this way, the number of correct predictions (correctly classified instances or CCI) could be calculated. In Figure 3 one can observe that artificial neural networks (ANNs) make better predictions than simple probabilistic guesses. The reliability of the models is the highest for very common (Chironomidae, Tubificidae) and extremely rare taxa (*Aplexa*, *Ephemera*, *Armiger*), but the added value of the artificial neural network compared to simple probabilistic guesses is the lowest under these circumstances.

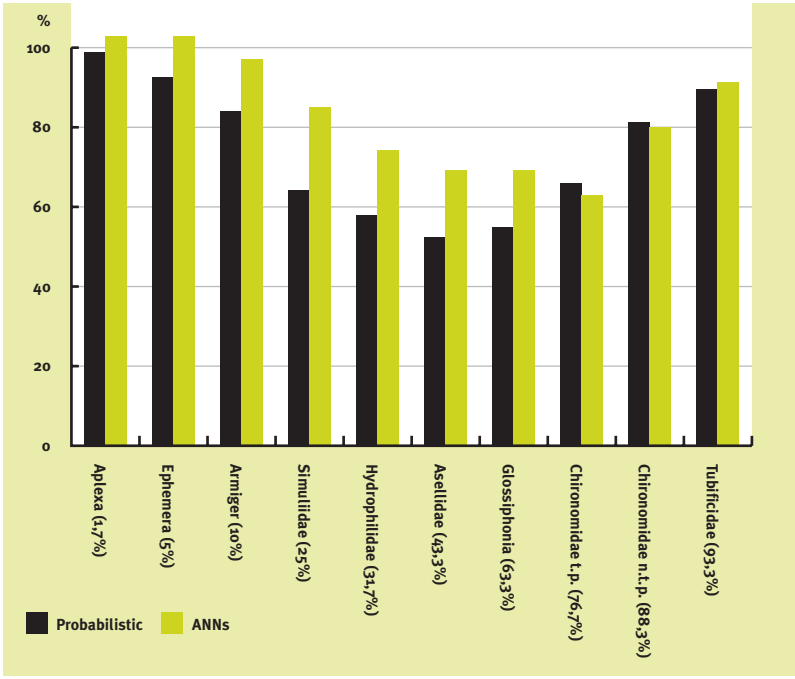


Figure 3
Prediction of macroinvertebrate taxa based on artificial neural networks (ANNs) compared to simple probabilistic guesses.

If the previously described modeling procedures, merely based on data mining techniques, are followed, models that reliably describe the processes of the focal ecosystem during the specific data acquisition period are attained. These models are not necessarily valid for any other period, because ecosystems tend to regulate, modify and change their parameters to respond to change in the prevailing conditions. These parameter changes are determined by the forcing functions and the interrelations between the state variables. Recently developed techniques, namely structural dynamic models [Jorgensen, 2001], therefore introduce parameters that can change according to forcing functions and general conditions for the state variables to continuously account for adapta-

tion. This type of model accounts for the change in species composition as well for the ability of species to alter their properties e.g. to adapt to the prevailing conditions imposed on their environment. Structural dynamic models have to be constructed mainly by means of expert knowledge, due to a lack of appropriate measurement sets. The so-called goal functions of these models describe the change of the parameters in function of changes imposed to the ecosystem. Exergy [Jorgensen, 1997] has been most widely used a goal function within this context. The development of these structural dynamic models is however only in an experimental phase and further research will be necessary to make these models applicable for river management.

Data visualization

Increasingly, geographical information systems (GISs) are emerging as tools for the identification and quantification of potential ecological impacts. Linked to comprehensive databases on the distributions of abiotic and biotic variables, they offer powerful techniques for addressing 'where?' and 'what if?' questions about the location and magnitude of interactions between ecosystem components and stress-generating activities [Treweek, 1999]. In particular, simple overlaying techniques (Figure 4) can be used to identify (and measure) areas where there is an overlap between an activity and stress measured by an important ecological receptor [Treweek, 1999]. In this way GISs entail a fast information flow from field measurements to the responsible managers. Information visualization of management plans and their (predicted) effects provides responsible managers with deeper insights into the ecosystem problems and solutions, and can thus result in more sustainable decisions on river restoration and protection.

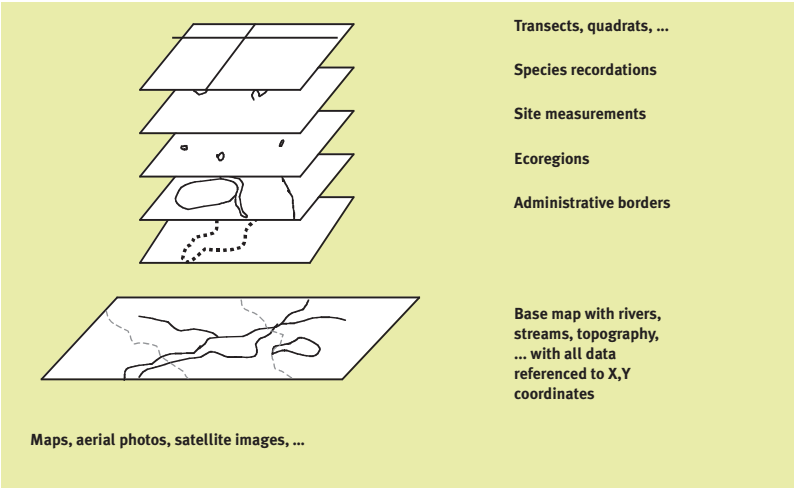


Figure 4
Overlay applications in a GIS (after Treweek, 1999).

Management applications of river ecosystem models

Models have several interesting applications in river management. They allow for a better interpretation of the results, easing the cause-allocation of the actual river status and increasing the insight needed to improve assessment systems (Figure 5). Models also allow for simulating the effect of potential management options and thus supporting decision making. The development of effective and efficient monitoring networks based on models is probably another important advantage [Goethals, 2001].

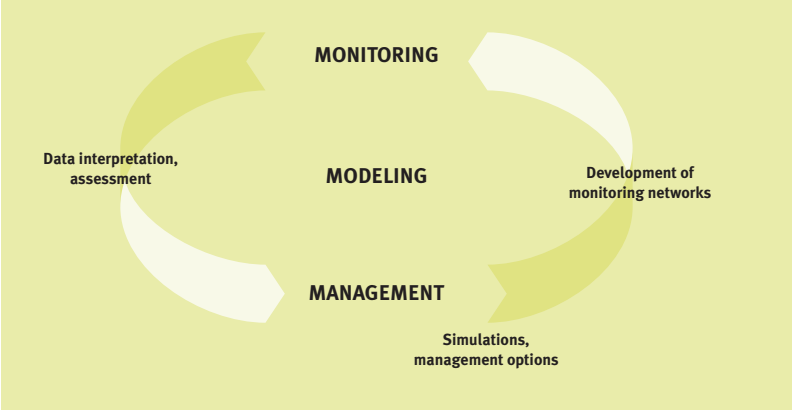


Figure 5
Potential applications of ecosystem models in integrated river management [Goethals, 2001].

Models make the quantification of relations in ecosystems possible and can in this way help to set standards for nature protection and restoration management. An example of quantifying relations between organisms and physical-chemical and structural conditions is presented in Case study 5. The convenience of model simulations to select restoration options for a disturbed aquatic ecosystem is shown in Case study 6.

Case study 5: ANN model simulations for analyses of macroinvertebrate habitat preferences

Sensitivity analyses can be used to acquire insight into the applied 'concepts' of ANN black-box models simulating habitat preferences of the macroinvertebrate taxa in the Zwalm river basin. Sensitivity analyses allowed the quantification of the impact of input variables on the presence or absence of macroinvertebrate taxa. In many cases, the ANN models detected a relevant relation between the input variables and the probability of presence of macroinvertebrate taxa. This provides some insight in the habitat preference of all taxa, which delivers substantial information for river ecosystem management. This can also be used to set environmental standards for nature protection and restoration. In Figure 6 and Figure 7, the impact of flow velocity and dissolved oxygen on the probability of presence of Gammaridae is shown. Figure 6 indicates that Gammaridae prefer faster running waters, while Figure 7 demonstrates that Gammaridae are

Figure 6
The impact of flow velocity on the probability of presence of Gammaridae.

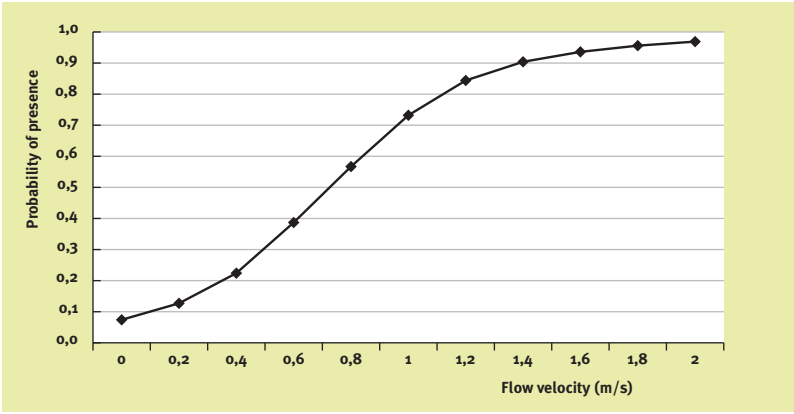
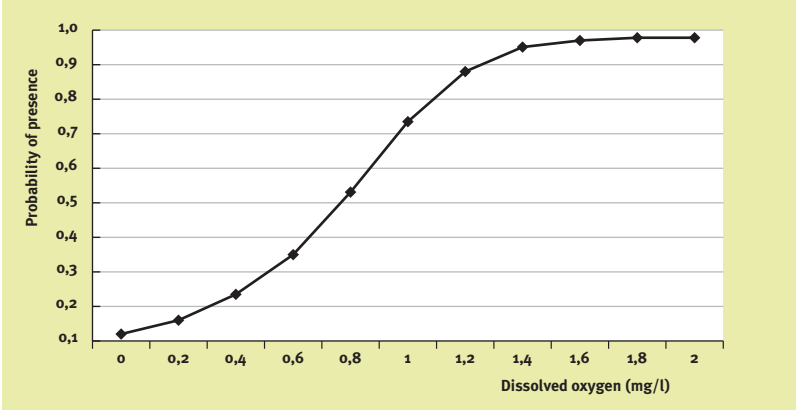


Figure 7
The impact of dissolved oxygen on the probability of presence of Gammaridae.



more abundant in well oxygenated waters, and are nearly always present when the oxygen level is higher than 7 mg/l.

Case study 6: Application of ANN models in river restoration management

An illustration about the application of ANN models to mitigate human impacts on the Bettelhovebeek is presented in this case study. Several structural modifications are affecting the biological communities at this site dramatically. The aim of this study was to determine the most efficient restoration option to obtain a stable biological ecosystem meeting the minimal river water quality standards for Flanders, such as the Belgian Biotic Index (BBI) equal or higher than seven. In Table 2, only predictions for the best restoration option are summarized, based on a selection made from a set of simulations with artificial neural network models [Dedecker, 2001]. The results indicate that after river restoration some macroinvertebrate taxa, indicative for a good water quality and not currently present, will colonize the site again. Also the predicted BBI changes from a moderate to a good quality, illustrating that the basic water quality standards for Flanders are met under the improved conditions.

	Before river restoration	After river restoration
Structural characteristics	**Actual bad situation**	**Future good status**
artificial embankments	concrete and iron	none
meandering	none	moderate
deep/shallow variation	none	moderate
hollow river beds	none	well-developed
Macroinvertebrate taxa	**Actual bad status (field measurement)**	**Future good status (predicted by ANN-model)**
Sialis	absent	present
Limnephilidae	absent	present
Simuliidae	absent	present
Belgian Biotic Index	**Moderate quality (BBI=5)**	**Good quality (BBI=7)**

Table 2
Optimal restoration option for the Bettelhovebeek and predicted macroinvertebrate taxa and BBI after river restoration.

Conclusions

Information stream optimization tools for ecological management, 'ecological informatics', are increasingly needed to support the decisions of environmental managers. Database management, development of predictive models and knowledge visualization are probably the main exponents of ecological informatics. Several case studies illustrated the development and use of these features in ecology, and showed the ease of practical application . Particularly those models that can offer predictions of faunal responses to changes in environmental conditions (e.g. changes in discharge regime, dissolved oxygen level, etc.) can be of considerable value for river management.

Acknowledgements

The authors would like to thank the Scientific Research Foundation of Flanders (FWO-Flanders) for its financial support (project 3G01.02.97) that made the research on ecological informatics on the Zwalm river possible.

References

– AMINAL. (1999). Control of Sediment Transport in Unnavigable Watercourses as Part of Integrated Water Management. Zwalm river basin project. Ghent, Belgium (in Dutch)
– Barros, L.C., R.C. Bassanezi, P.A. Tonelli. (2000). Fuzzy Modelling in Population Dynamics. Ecological Modelling **128**:27-33
– Breimann, L., J.H. Friedman, R.A. Olshen, C.J. Stone. (1984). Classification and Regression Trees. Pacific Grove, Wadsworth
– Caldarelli, G., P.G. Higgs, A.J. McKane. (1998). Modelling Coevolution in Multispecies Communities. Journal of Theoretical Biology **193**:345-358

- Carchon, P., N. De Pauw. (1997). Development of a Methodology for the Assessment of Surface Waters. Study by order of the Flemish Environmental Agency (VMM). Ghent University, Laboratory of Environmental Toxicology and Aquatic Ecology, Gent, Belgium (in Dutch)
- Dedecker, A. (2001). Modelling of Macroinvertebrate Communities in the Zwalm River Basin by Means of Artificial Neural Networks. M. Eng. Thesis. Ghent University, Faculty of Applied Biological Sciences, Gent, Belgium
- Demuth, H., M. Beale. (1998). Neural Network Toolbox for Use with MATLAB. User's Guide. Version 3.0. The Mathworks, Natick, USA
- Goethals, P., N. De Pauw. (2001). Development of a Concept for Integrated Ecological River Assessment in Flanders, Belgium. Journal of Limnology (accepted)
- Gronewold, A., M. Sonnenschein. (1998). Event-Based Modelling of Ecological Systems with Asynchronous Cellular Automata. Ecological Modelling **108**:37-52
- Jorgensen, S.E. (1997). Integration of Ecosystem Theories: a Pattern. Second Revised Edition. Kluwer Academic Publishers, Dordrecht
- Jorgensen, S.E. (2001). Recent Trends in the Development of Ecological Models Applied on Aquatic Ecosystems. Ecological Modelling (accepted)
- Lauryssen, F., F. Tack, M. Verloo. (1994). Nitrogen Transport in the Zwalm River Basin. Water **75**:46-49 (in Dutch)
- Lek, S., J.F. Guegan. (1999). Artificial Neural Networks as a Tool in Ecological Modelling, an Introduction. Ecological Modelling **120**:65-73
- Michener, W.K., J.W. Brunt. (eds.). (2000). Ecological data: Design, Management and Processing. Blackwell Science, Oxford, UK
- NBN. (1984). Biological Water Quality: Determination of the Biotic Index Based on Aquatic Macroinvertebrates. NBN T92-402. Institut Belge de Normalisation (IBN), Belgium (in Dutch and French)
- Pauw, N. De, G. Vanhooren. (1983). Method for Biological Assessment of Watercourses in Belgium. Hydrobiologia **100**:153-168
- Pauw, N. De, R. Vannevel. (1993). Macroinvertebrates and Water Quality. Dossiers Stichting Leefmilieu 11. Stichting Leefmilieu, Antwerp, Belgium (in Dutch)
- Rumelhart, D.E., G.E. Hinton, R.J. Williams. (1986). Learning Representations by Back-Propagating Errors. Nature **323**:533-536
- Soresma. (2000). Environmental Impact Assessment Report on the Development of Fish Migration Channels and Natural Overflow Systems in the Zwalm River Basin. Soresma Advies- en Ingenieursbureau, Antwerp, Belgium (in Dutch)
- Treweek, J. (1999). Ecological Impact Assessment. Blackwell Science, Oxford, UK
- VMM. (2000). Water Quality - Water Discharges 1999. Flemish Environmental Agency, VMM, Erembodemgem, Belgium (in Dutch)

2.3.10 Data mining for natural language processing

Antal van den Bosch[1]

The field of computational linguistics, an interdisciplinary field between linguistics and artificial intelligence, is concerned with developing models and systems for the processing of natural language. In terms of applications, the field aims to produce systems that can assist people in writing, understanding, and translating language. Natural language processing systems are nowadays used as hidden modules such as style and spelling checkers in word processors and in search engines, but also as the virtual ears and mouths of automatic dialogue systems.

Data mining methods have been used in linguistics and subfields for a long time, inducing rules from relevant observations. They are used intensively in present-day research in natural language processing. In this section we illustrate that this has not been a smooth historical process: empiricism has returned to (computational) linguistics in the nineties, after an absence of decades. We give a brief review of present-day empirical linguistics, and conclude that this area can be expected not only to rely heavily on data mining techniques in the future, but also to continue to be a source of new developments in data mining.

A brief history

A prime reason for using natural language is to convey information. As most language users know, albeit implicitly, this information is sometimes obviously present, but is mostly hidden in the surface message. Speakers and writers pack information in utterances that obey a language-dependent word ordering (syntactical) and word formatting (morphological) system. This system, usually evolved through centuries and constantly changing, can be explained to some degree by rules and exceptions, but it is always hard to learn it in detail (vocabulary, idioms, irregular past tenses, case systems, etc.) in order to cope with the inherent ambiguity of natural language at an advanced level.
Nevertheless, driven by the impact of the Cartesian linguistic theories of Noam Chomsky, linguistics and its applied subfield of computational linguistics have worked for three decades with the hypothesis that natural language use and processing could be modeled by a system that maps utterances to logical formulas [Chomsky, 1957] using a fundamentally rule-based formalism. Although this school has always met with some resistance (and indeed it has by far not met the goals it has set on e.g. machine translation and message understanding), it was only during the past decade (the 1990s) that a counter-hypothesis was voiced and, with increasing success, subsequently tested: that language

1 Dr A. van den Bosch,
Antal.vdnBosch@kub.nl,
ILK / Computational Linguistics,
Tilburg University, The Netherlands.
http://ilk.kub.nl/~antalb/

could be modeled by a system that would be purely databased, i.e. empirical or statistical. In a computational guise, this new stream in computational linguistics recaptures the basic claims of pre-Chomskyan linguists such as Bloomfield and Firth that no theory of language use and processing would be valid or usable unless it was grounded in data [Firth, 1964].

Empirical natural language processing

The core idea of empirical natural language modeling is that language can be processed by analogy to stored instances of language, or probabilistic models thereof [Manning, 1999].

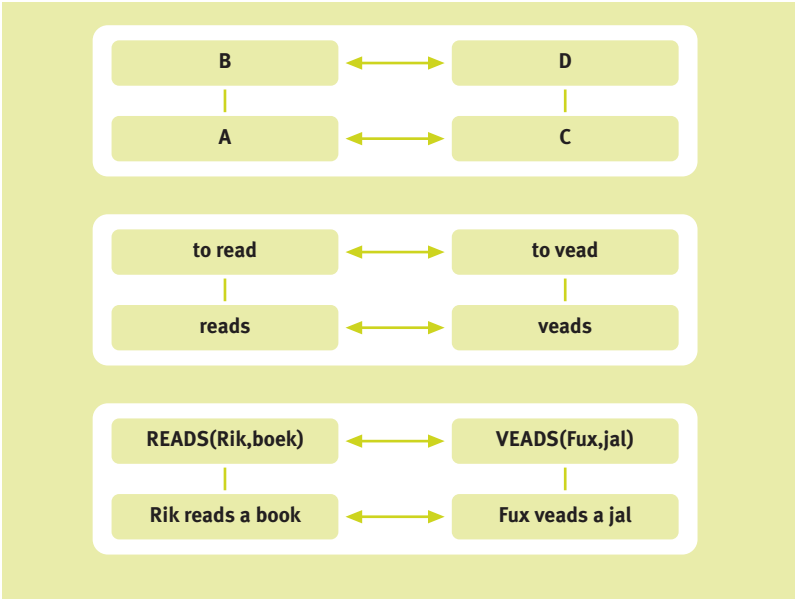


Figure 1
Analogy in language. The analogy principle says that 'A:C=B:D' (top); if A and C stand in a similarity relation, then B and D stand in the same similarity relation. In the middle, this principle is used to predict that the infinitive form of the nonsense verb form 'veads' is 'to vead', in analogy with 'reads'/'to read'. The bottom example assigns a simple meaning to the nonsense sentence 'Fux veads a jal' ('VEADS' being a predicate with arguments 'Fux' and 'jal'), in analogy with the sentence 'Rik reads a book'.

Similar utterances have similar meanings — given that an appropriate definition of 'similarity' can be devised. As most of the current work in this area shows, similarity expressed at very simple and local levels can yield very accurate models of morphology and syntax, and in this sense data mining, as an overlapping name for the statistical and machine learning methods used in the field thus far, has already supplied the field with valuable techniques. After morphology and syntax comes semantics — the characterization of the meaning or intention of utterances. As it was to Chomskyan linguistics, modeling semantics is a challenge also to empirical natural language modeling, and data mining is expected to play a significant role in this endeavor the next decade. The utility of being able to handle semantics on the other hand is enormous. At an increasing pace, people are searching information in large databases. Keyword search is the inefficient common method, and it is becoming clear that retrieving results on information queries could be much improved, if computational lin-

guistic systems existed that could extract relevant information from texts, classify them in categories, and make abstracts from them. Building tools for information extraction from texts, or text mining, is actually the most-studied topic in world-wide computational linguistics in 2000, and it can be expected to change the field dramatically in the next decade. Once the scientific endeavor has produced applications that can be integrated in home computer software (word processors and browsers), text mining tools can be expected to be commonplace within 15 years, assisting individuals, including scientists (see Paragraph 5.4), in gathering their information through the day.

Current methods and data

The computer revolution has dictated that if text is produced, it is often digitized – it is increasingly common for text to be available *only* in digital form. Also, previously written and printed material (books, laws) is being digitized through OCR techniques, stored, and to some extent made accessible. With hardly any effort from the scientific field, enormous amounts of textual data can be harvested by just tapping electronic databases — within the limits of copyrights and privacy laws. The World Wide Web, for example, is often called the largest open text corpus in the world [cf. Lawrence, 1999]. On the other hand, there are relatively few high-quality text collections such as modern literature or newspaper archives in the public domain. Specific semi-commercial institutions such as the Linguistic Data Consortium (LDC) [2] and the European Language Resource Association (ELRA) [3] have been set up as collectors and developers of standardized, high-quality linguistic data for the scientific community. Annotation standards have been proposed with an increasing sophistication during the 1990s, and are now seamlessly integrated with the global standardization developments in markup languages (XML) [4]. More on XML in Section 5.6.1 Web mining, and on the CD [5].

Increasing effort is invested by institutions such as LDC and ELRA, but also by many research groups into the annotation of raw data at specific linguistic levels, for example the phonetic transcription of speech, syntactic analyses of written sentences, marking specific types of information in text such as names or important words, classifying turns in dialogues as questions, answers, or fillers, etc.

This usually demands the investment of time and the training of skilled human annotators, and it is often criticized as posing a knowledge acquisition bottleneck that can be as serious as the traditional bottleneck involved in producing rules by expert introspection. On the other hand, creative solutions are developed in using data mining for the semi-automatic annotation of language data [Brants, 2000], and in using unsupervised data mining methods to discover annotation classes automatically [Kehler, 1999].

2 Web page of LDC: http://www.ldc.upenn.edu

3 Web page of ELRA: http://www.icp.grenet.fr/ELRA/

4 XML portal for industry: http://www.xml.org

5 CD-rom: ..\papers\XML in 10 points.htm , ..\papers\Extensible Markup Language (XML) 1_0 (Second Edition).htm

```
((S
  (NP-SUBJ
    (NP (NNP Pierre) (NNP Vinken))
    (,,)
    (ADJP
      (NP (CD 61) (NNS years)
      (JJ old)
    (,,))
  (VP (MD will)
    (VP (VB join)
      (NP (DT the) (NN board))
      (PP-CLR (IN as)
        (NP (DT a) (JJ non-executive) (NN director) ))
      (NP-TMP (NNP Nov.) (CD 29) )))
```

Figure 2
Full syntactic analysis of the sentence 'Pierre Vinken, 61 years old, will join the board as a non-executive director Nov. 29.'; the first sentence of the LDC Wall Street Journal Penn Treebank corpus, using brackets and indentations, all syntactic roles and relations (e.g. subject phrase, NP-SUBJ; temporal phrase, NP-TMP) within the underlying syntactic tree are represented.

State of the art

Although there is an enormous demand from the market , the field has not yet made it into the average desktop computer. Speech recognition, dialogue systems, authoring tools, and linguistically clever information retrieval and extraction have not arrived at a level of global use, and this is partly because the accuracy of all of these different high-end applications is still too low for this type of usage. This is the basic reason that the state of the art in computational linguistics and related disciplines can be best found in competitions organized within the field itself. In North America, the yearly MUC (the Message Understanding Competition) and TREC (the Text Retrieval Competition[6]) draw the best research teams from both universities and companies to compete with each other on the most accurate summarization, information retrieval, or question answering. Another example is SENSEVAL[7], a recurring competition between systems that perform word sense disambiguation in free text.

In general, the best known systems operate in relatively limited real-world domains such as booking and travel scheduling. In these domains language use is generally more regularly due to the use of jargon and conventions. Furthermore, the modality of domain language use is often constrained (utterances are made e.g. only by telephone, only in monologues, in man-machine dialogues, etc.). Perhaps the most illustrative example of a big project that was relatively successful in its highly ambitious goal was the VERBMOBIL project[8] [Wahlster, 1997], which aimed at developing a speech-to-speech translation system between English (or German) and Japanese, on the domain of the scheduling of appointments of two business people over the phone.

Prospects and challenges

As more language data becomes available, both raw and annotated, and as technology produces faster computers and greater storage capacities, in desktop computers also, the field of computational linguistics will develop increasingly quickly. It will also produce applications that will enter the consumer mar-

6 Web page of TREC: http://trec.nist.gov

7 Web page of SENSEVAL: http://www.itri.brighton.ac.uk/events/senseval/

8 Web page of VERBMOBIL: http://verbmobil.dfki.de

ket on a much wider scale than at present. It will help guide the information streams on the Internet to become much more efficient and precise; a much-needed improvement.

It is absolutely clear that data mining will be essential in this development. The challenges in computational linguistics are typical data mining questions, and will need data mining solutions: what are the best representations of language, besides the raw digital material that can be used in high-end applications? Feature representation, selection, and combination is a core problem in data mining and language with its many thousands and millions of words, sentences, and texts offers a very critical challenge to the scaling abilities of data mining methods.

Given this, it is not surprising that some new data mining methods have emerged from the computational linguistic field. Inductive Logic Programming finds part of its basis in the field [Muggleton, 1997]. Some significant developments in Memory-Based (Instance-Based) Learning have come from applications in natural language processing [Daelemans, 1999]. Research in text categorization has boosted the interest in Support Vector Machines [Joachims, 1998]. During the last two decades, many statistical methods for modeling, classification, and feature selection have been relatively widely tested in the natural language domain (to name but a few, artificial neural networks (see 6.2.8 Neural networks), Bayesian classifiers (see 6.2.9, Naïve Bayes classifier), Maximum Entropy modeling, Hidden Markov models, (see 6.2.10, Hidden Markov Models)). It has often been stated that many of these developments will prove (and some already have proven) to be applicable to other domains in which data mining is used for discovering knowledge in sequences, such as in DNA strings or in time sequences such as financial indicators.

References

- Brants, T., O. Plaehn. (2000). Interactive Corpus Annotation. In Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000). Athens, Greece
- Chomsky, N. (1957). Syntactic Structures. Mouton, Den Haag
- Daelemans, W., A. van den Bosch, J. Zavrel. (1999). Forgetting Exceptions is Harmful in Language Learning. Machine Learning **34**:11-43
- Firth, J.R. (1964). Tongues of Man and Speech. London
- Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Machine Learning ECML'98. Lecture Notes in Artificial Intelligence 1398. Springer Verlag, Berlin
- Kehler, A., A. Stolcke. (eds). (1999). Proceedings of the ACL'99 Workshop on Unsupervised Learning in Natural Language Processing. ACL, New York

– Lawrence, S., L. Giles. (1999). Searching the Web: General and Scientific Information Access. IEEE Communications **37** (1):116
– Manning, C., H. Schütze. (1999). Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, MA
– Muggleton, S. (ed.). (1997). Inductive Logic Programming: Selected Papers from the 6th International Workshop. Springer Verlag, Berlin
– Wahlster, W. (1997). VERBMOBIL: Erkennung, Analyse, Transfer, Generierung und Synthese von Spontansprache. DFKI Gmbh. Verbmobil Report 198

2.3.11 Application of data mining tools in the behavioral sciences

Elise Dusseldorp[1], *Jacqueline Meulman*[2]

Introduction

The behavioral sciences encompass psychology as well as the educational sciences. The origin of psychology as a science is often equated with the origin of physical psychology at the end of the 19th century. In 1879, Wundt founded the first psychological laboratory in Europe, in Leipzig. In this laboratory, the *general human functions* (e.g. auditory and visual perception) were studied, and measurement instruments were developed.

In this Section, however, we will focus on another direction in the behavioral sciences, that is, the study of *individual differences* in human functioning. The book 'Hereditary Genius' by Francis Galton (1869) may be considered as part of the dawn of the study of individual differences, and the beginning of the development of the mental test. This movement had a huge influence on the development of the behavioral sciences in the United States of America.

Galton developed statistical techniques to analyze data on individual differences and, thus, influenced the development of statistics for behavioral scientists (psychometricians) and biometricians. For example, he invented regression analysis for analyzing measurements of height (stature) from children and their parents, and described the regression-to-the-mean phenomenon (1886). Furthermore, he adapted the normal distribution of Quetelet, but contrary to Quetelet, who considered variation in scores as measurement errors, Galton considered variation as essential.

Cattell (who was inspired by Galton) is assumed to have introduced the term mental test for the first time in the behavioral sciences in his article 'Mental tests and measurements' (1890). He advised Thorndike to apply his animal intelligence techniques to children and adolescents, which contributed to the development and application of mental tests in schools and education. In his book 'Educational Psychology' (1913) Thorndike describes tests developed to predict school success. Achievement tests were born, and used to set norms (standards) that had to be achieved by children of a certain age. Children who could not achieve a norm were considered to have learning deficiencies.

The idea behind a mental test is that people with different scores on the test, differ in an immeasurable latent property (e.g. intelligence). One of the motives underlying the endeavors of Galton and Cattell was to select individuals with higher intelligence and to stimulate marriages among them, in this way 'improving' the human race in the long-term. Others, fortunately, had more socially-oriented motives. Binet, for example, developed tests to select children with learning deficiencies, and was an advocate of special classes for this group of children.

1 Dr E.M.L. Dusseldorp, dusseldorp@fsw.leidenuniv.nl, Data Theory Group, Faculty of Social and Behavioral Sciences, Leiden University, The Netherlands.

2 Prof Dr J.J. Meulman, meulman@fsw.leidenuniv.nl, Data Theory Group, Faculty of Social and Behavioral Sciences, Leiden University, The Netherlands.

A related issue in this area of research is the assumed contrast between the role of nature and nurture in behavioral science theories. Scientists who emphasized nature (e.g. Galton and Cattell) believed that differences in functioning between individuals were mainly determined by heredity, while scientists who emphasized nurture (e.g. Skinner and Pavlov), believed that these differences were determined by education and environmental influences. Modern approaches in psychology and education consider human functioning (e.g. learning) as a dynamic process and put more emphasis on the interaction between human and environment.

Nowadays, psychological achievement and screening tests are used in all kinds of situations (e.g. job selection, or diagnosis of patients). A wide variety of tests exist, ranging from general to more specific ones. The latter focus on specific aspects of human functioning for particular groups of individuals. From a historical perspective, a *psychological test* is mostly associated with an intelligence test. Therefore, we prefer to speak of psychological and educational *measurement instruments*. We will show in this Section an application of various data mining techniques to data collected with a psychological measurement instrument, a questionnaire. This questionnaire has been developed to measure differences between coronary heart disease patients with regard to unhealthy behavior, psychological resources, social support, and quality of life [Elderen, 1997].

Specifics of the behavioral sciences

Research in the behavioral sciences is directed at discovering more knowledge about human functioning and behavior, in its broadest sense. In general, psychology focuses more on the behavior of the individual, while education focuses on the influence of environment (upbringing, education) on behavior.

Examples of research questions are: What causes a specific behavior? What factors influence behavior? Why do people change their behavior? Can we predict behavior (change)? What is the effect of a treatment (e.g. therapy) on behavior?

In the classical empirical process needed to answer these questions, several steps can be distinguished.

- Firstly, the development of theory and the construction of hypotheses. On the basis of *a priori* knowledge, that is, results from previous research, the researcher formulates new hypotheses for a study. Note: in Section 2.2.4, a method is outlined that uses data mining to generate hypotheses from available data.
- Secondly, the design and planning of the study; for example, a longitudinal design (with two or more measurement points) or a cross-sectional design (one measurement point for different cohorts).
- Thirdly, making psychological concepts operational and selecting measurement instruments. For example, the researcher may choose to use an existing questionnaire to measure 'anxiety' or to develop a new instrument.

– In the fourth step, the data are collected. In general, the researcher can not use the whole population in a study, but only a sample; this sample must be representative for the population of interest. Also, the sample size has to be large enough to draw useful and statistically sound conclusions.
– In the fifth step, the data are analyzed; the researcher may choose from many (multivariate) analysis techniques developed for behavioral science data (see Sections 6.2.2 to 6.2.5: Regression, Discriminant analysis, Multidimensional scaling).
– In a last step, the results are reported, mostly in a scientific article.

These six steps describe the empirical process globally. In the practice of behavioral science, the content of these steps may vary. One important variation has to be explained. Sometimes a research project involves a new subject, and explores new areas of research. The formulation of hypotheses on the basis of a priori knowledge is not possible in these cases, simply because results of previous research are not available. In such situations, the steps of the empirical process will have a more exploratory nature. In the first step, the research questions will be formulated as concisely as possible. Generally, in the third step, several different measurement instruments (existing ones and newly developed ones) will be used, and many variables will be investigated. In the fifth step the data analysis will be directed mainly at generating hypotheses for future research.

Characteristics of the available data

Most data collected in the behavioral sciences are measurements (e.g. personality characteristics or observable behaviors) of persons (e.g. patients, children, or clients). The measurements are called variables, and are categorical (discrete) or continuous. The distinction between categorical and continuous is related to the scaling level of the variables: categorical variables are usually given a nominal or ordinal scaling level, and continuous variables a numerical one. A nominal scaling level divides persons into distinct, unordered categories, for example: male or female. For an ordinal scaling level the order of the categories is important, for example: never, sometimes, often, and always. For a numerical scaling level the interval between different category values is also important, for example: 1, 2, 3, or 4 years. In other words, we can say that 4 years is twice as much as 2 years, but we can not say that *always* is twice as much as *sometimes*.
The majority of behavioral science data is collected by self-report questionnaires. This implies that in contrast to other disciplines where automated data collection is the rule, in the behavioral sciences data are sparse and hard to obtain. Typically, several items in a questionnaire are used to measure one construct (e.g. 'anxiety'). The scores on the items belonging to one construct are added up to obtain a sum score on a scale. To estimate the reliability of a scale, Cronbach's alpha is often used [Cronbach, 1951]. A scale with a Cronbach's

alpha of .80 or higher can be considered a reliable scale, indicating that the test results are consistent.

The validity of behavioral data is often a point of discussion, because it is very difficult to assess. If one has a clear criterion (a 'golden standard') about what property a scale is aimed to measure, then the validity can be easily determined. Such a golden standard is mostly lacking, however, and the validity of a scale is determined by comparing (correlating) the scales with scales obtained from other measurement instruments that measure *approximately* the same property (for an example see [Hillers, 1994]).

Mostly, the owner(s) of the data is (are) the researcher(s) who carried out the research project. When the results of a study are reported in a journal article, the author(s) have to comply with the ethical principles of the journal. Many journals in the behavioral sciences have adopted the ethical principles of the American Psychological Association (APA); examples of such journals are Psychological Bulletin, the Journal of Consulting and Clinical Psychology, Child Development, and Psychological Methods. One of the APA principles is about sharing data: "After research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use such data only for that purpose (..)" [APA, 1994, p. 293]. Authors in APA journals are expected to retain their data for a minimum of 5 years after their article has been published.

The four most important (non APA) behavioral science journals devoted to the development of new data analysis methods are Psychometrika, the Journal of Classification, the British Journal of Mathematical and Statistical Psychology, and the Journal of Educational and Behavioral Statistics.

State of the art for data mining applications

If we consider data mining in its broad sense, and refer to the Knowledge Discovery in Databases process (KDD), we distinguish five steps [Gaul, 1999] see Chapter 6.1:

1 task analysis
2 preprocessing
3 data mining
4 postprocessing (reporting)
5 deployment.

Whereas in the empirical scientific process three steps are related to the data *collection* process (second, third, and fourth step), in the KDD process the data have often been collected before the process starts [Hand, 2000, p. 112]. There is, however, much overlap between the data analysis step in the scientific process, and the preprocessing and data mining steps of the KDD process.

Preprocessing

The preprocessing of data in the KDD process involves missing data analysis, transformation of data (standardization, scale transformation, and construction of variables), and feature selection (also called 'variable selection' or 'dimension reduction'). Behavioral science data often consists of *many* measurements on relatively *few* persons (rarely exceeding 1,000, often between 40 and 400). The reliability of most statistical models, however, decreases, when too many variables are included relative to the number of persons. Generally, for a reliable regression model about 15 cases per predictor are considered necessary in the behavioral sciences [Stevens, 1992]. Consequently, an important application field for data mining in the behavioral sciences is *selection of variables*. Four different methods will be used to select variables for the analysis. Among these are two commonly used methods (one from statistics and one from machine learning) and two state of the art multivariate data analysis methods. Since selection can be considered the most important step in this situation, we will focus on this and use logistic regression as a default for the data mining part.

Suitable techniques for the selection of variables

Stepwise selection

Basically a statistical method, stepwise selection in regression analysis is mostly used in the behavioral sciences to reduce the number of (predictor) variables in regression. Two types of stepwise selection exist: forward or backward selection. In the example we will use backward stepwise selection. An advantage of backward elimination of predictors over forward inclusion is the sensitivity to suppressor effects. The latter term refers to the situation, when a variable appears to have a statistically significant effect only, when another variable is controlled or held constant [Menard, 1995].

Decision tree analysis

Decision tree analysis originates from machine learning (also referred to as classification tree or rule induction method, CART for short; see 6.2.7 Classification, Decision tree learning). If one is not sure which variables are important, a recommended strategy in the field of data mining is to use decision trees as a first step of analysis to select variables [Clementine, 1998]. As a second step a neural network may be trained, or logistic regression may be applied.

Multiple additive regression trees

In MART [Friedman, 1999; Friedman, 2000] the response variable may be categorical or numerical. The predictors are treated as categorical or ordinal (as in a decision tree analysis). MART incorporates the idea of *boosting*: a general method of producing a very accurate prediction rule (a *strong learner*) by com-

bining rough and moderately inaccurate rules (*weak learners)* [Freund, 1997]. In MART several weak learners (i.e. regression trees) are created by a *stagewise* approach of iteratively fitting the residuals of a tree. In other words, several regression (or decision) trees are fitted consecutively, and each new tree is fitted on the residuals of the previous trees. The total number of trees being fitted equals the number of so-called boosts. The trees are combined into a strong learner by taking a weighted sum. The weight of each subsequent tree is estimated using a steepest-descent method. In this way, the predictive accuracy of MART is in most cases much higher than that of one single regression or decision tree. However, the interpretation of a MART solution is more difficult.

Categorical principal component analysis

CATPCA is a very general data (or dimension) reduction method [Meulman, 1999]. It is a generalization of principal component analysis, suitable for the analysis of categorical and ordinal data. The variables to be analyzed by CATPCA can be scaled at three different scaling levels (nominal, ordinal or numerical), for each variable separately. CATPCA scales the data optimally, according to the scaling level chosen, hence the name optimal scaling technique. The transformed variables can be saved, and used for further analysis.

A number of differences between the four methods have to be mentioned. First, in backward stepwise selection, the number of predictors is reduced *during* the logistic regression analysis. The other three methods can be used to reduce the number of predictors *before* the regression analysis is applied. Second, in stepwise regression, CART and MART, the relationship between the predictor variables and the response is explicitly maximized. In a standard CATPCA, the interdependence of predictors that measure the same property is maximized. However, by including the response variable in the analysis with a larger weight, the interdependence is maximized in relation to the response variable. Third, in CATPCA and MART all available cases are used in the predictor reduction process, while in the stepwise procedure cases with one or more missing values on the predictors are deleted from the analysis[3].

An example with real life data

We will illustrate the preprocessing and data mining steps in the behavioral sciences with a real life example. In this example, data from the field of health psychology are used, and refer to a psycho-educational prevention study [Elderen, 2001, in press]. This study has an explorative nature, and incorporates many variables. Regardless of the preprocessing method used, logistic regression will be used as the data mining technique.

The Psycho-Educational Prevention (PEP) study focuses, among others, on the research questions: Which patients quit smoking after a cardiac event, and

3 In decision tree analysis some implementations (e.g. CART) use all available cases and some implementations (e.g. the 'tree' function in Splus) delete cases with missing values on the predictors.

Domain	Predictors	Categories
demographic	gender	male/female
	work	no/yes (i.e. more than 10 hours a week outside the house)
	education	ranging from 1 = only elementary school, to 4 = university, high vocational school, or high school
	living situation	with partner/without partner
	age	ranging from 34 to 70 years
medical	cardiac event	1 = first MI; 2 = first PTCA; 3 = first CABG; 4 = first combination of events; 5 = recurrent event
	NYHA	ranging from 0 = no chest pain to 4 = no activity without chest pain
	duration	duration of cardiac complaints in months, ranging from 0-288
psychological	self-efficacy1	general self-efficacy; ranging from 10 (little) to 40 (very much)
	self-efficacy2	appraisal of self-competence in coping with cardiac event: 1 = not, to 4 = totally competent
	self-efficacy3	need of counseling in 'working through' the cardiac event: no/yes
	self-efficacy4	specific self-efficacy in quitting smoking, ranging from 1 (little) to 5 (very much)
	self-efficacy5	need of guidance to quit smoking: yes/no
	coping1	approach behavior, ranging from 12 (little) to 36 (very much)
	coping2	self-blame, ranging from 5 (little) to 19 (very much)
	coping3	avoidance, ranging from 20 (little) to 52 (very much)
social	quantity1	number of persons who give emotional support, ranging from 0 to 8
	quantity2	number of persons who give instrumental support, ranging from 0 to 8
	quality1	general satisfaction with social support: 1 = not; 2 = partly; 3 = very satisfied
	quality2	satisfaction with support of partner: 1 = not 2 = partly; 3 = very satisfied
	quality3	satisfaction with family support, ranging from 29 (little) to 65 (very much)
quality of life (QOL)	atypical physical complaints	summation of complaints of shortness of breath, chest pain, physical and mental fatigue, ranging from 0 to 6.
	emotional QOL	ranging from 15 (low) to 77 (high)
	physical QOL	ranging from 13 to 70
	social QOL	ranging from 18 to 49
	positive feelings	ranging from 4 to 16
treatment	cardiac rehabilitation	1 = FIT; 2 = FIT + INFO; 3 = FIT + INFO + PEP
	hospital	1 = Tilburg; 2 = Zwolle; 3 = Eindhoven

Table 1
Overview of predictors in the psycho-educational prevention study. MI = myocardial infarction; PTCA = percutaneous transluminal coronary angioplasty; CABG = coronary artery bypass grafting; FIT = physical training program; INFO = health education program; PEP = psycho-educational prevention program.

which type of cardiac rehabilitation stimulates patients to quit smoking? To answer these questions, patients who had experienced different cardiac events (see Table 1) and who stayed in three hospitals in the Netherlands, participated in the study. They were randomly assigned to three cardiac rehabilitation programs. Patients were offered either physical training sessions (FIT), or information sessions (INFO) in addition to FIT, or psycho-educational prevention sessions (PEP; based on rational emotive therapy) in addition to INFO and FIT. The patients received a questionnaire [Elderen, 1997] at two time points: before the cardiac rehabilitation (M1), and after the rehabilitation (M2). In the present study, we included in the analyses only patients who smoked before the cardiac event (the sample size (N) = 173, which was 51% of the total sample). To investigate the research questions, 29 variables were measured. We will refer to one variable as the *response* (i.e. smoking behavior at M2), and to the other 28 variables as *predictors*. A health psychologist divided the predictors into six domains beforehand on the basis of their conceptual relationships: demographic, medical, psychological, social, quality of life, and treatment. All predictors were measured at M1. Table 1 displays an overview of the variables in each domain.

As we wished to predict which patients quit smoking at M2 (measured by a binary, yes or no, response) we chose logistic regression as the method for analysis (after the preprocessing step). Logistic regression is especially suitable for prediction problems, when the response has only two categories. We have 173 patients and 28 predictors available for the analysis. Since about 15 cases per predictor are needed for a reliable regression equation, we first had to reduce the number of predictors to a maximum of 11. Table 2 shows the results of the selection of the predictors by the four different preprocessing methods.

Table 2
Overview of selected predictors by the four data mining methods. The predictors are ordered by frequency of selection from four to one.

Predictors	Stepwise	CART	MART	CATPCA
self-efficacy4	*	*	*	*
self-efficacy5	*		*	*
quantity1	*			*
social QOL	*			*
NYHA	*			
quality1	*			
quality2	*			
hospital	*			
emotional QOL			*	
work				*
cardiac event				*
cardiac rehabilitation				*

We used two measures to evaluate the accuracy of the classification rule estimated by each method. The first measure was the classification rate, that is, the percentage of subjects classified correctly according to the logistic regression model. To compute the classification rate, we assumed equal a priori class probabilities (i.e. priors); thus we assumed that the probabilities of quitting smoking and continuing smoking were equal in the population. Consequently, the classification threshold was set at 0.5. The value of this threshold, however, depends on both the priors and the costs of misclassification. Because we did not know these values in our population, we used a second measure, that is, the area under the ROC curve, called '*AUC*' [Hand, 2001]. The *AUC* is closely related to the *Gini* coefficient, and a more appropriate goodness-of-fit measure in case of unknown priors and costs. The lowest possible *AUC* is 0.5 and the highest is 1. The classification rate and the *AUC* were determined on the total sample (without cases with missing values on the selected predictors) and by the use of 10-fold cross validation (see Insert). To reduce the risk of over fitting, the 10-fold cross validation was performed on the *entire* process, that is, both the variable selection process and the logistic regression analysis were performed on each training sample. Subsequently, the classification rate and *AUC* were estimated on each test sample.

Insert: cross validation

Cross validation procedures determine the classification rate of a model on a test set. Cases in a test set are not used to estimate the parameters of a model (and also not used to select the variables for a model). In 10-fold cross validation, the total sample is randomly divided into 10 sets. Each time one set plays the role of the test set, and the other sets are the training set. This process is repeated, until all sets have entered as the test set, and thus all cases have been classified on the basis of the training set.

Stepwise selection

All 28 predictors were entered in the analysis. We used indicator coding for the categorical predictors. The backward stepwise selection procedure selected 8 variables in the final model. The number of patients used in the analysis was 108 (38% of the available patients were deleted due to missing data), and 80,6% were correctly classified. The *AUC* for the total sample was 0.879. In the 10-fold cross validation, the percentage correctly classified dropped to 69,4%, and the *AUC* to 0.610.

Decision tree analysis

Here, again, all 28 variables were entered in the analysis, and we indicated which variable was categorical or ordinal. The tree analysis automatically selects predictor variables during the tree-growing process. We applied the

tree-growing procedure recommended by [Breiman, 1984]: "grow a very large tree with a liberal stopping rule and then prune it upward." As stopping rule, we used a minimal terminal node size of 5. The value of the pruning parameter was determined by the use of cross validation (the function 'cv tree' in Splus). The variables selected in the pruned tree were chosen for the logistic regression analysis.
The pruning procedure in the decision tree analysis indicated that the best tree has two terminal nodes. The pruned tree on the total sample selected specific self-efficacy only. The number of patients used in the analysis was 108 (38% of the available patients were deleted), and 73,1% were correctly classified. The *AUC* was 0.794. This percentage correctly classified decreased to 69,4% in the 10-fold cross validation, and the *AUC* to 0.647.

Multiple additive regression trees

With MART we also entered all 28 predictors, indicating which were categorical or ordinal, and we inspected the misclassification risk. We varied the tree size, and chose the size having the lowest classification error (using the total sample). This turned out to be a tree size of 2 terminal nodes. The number of subjects in the test set was fixed at 10% of the sample size. The number of iterations was set at 200, the sample fraction at 1, and the learn rate at .025 (these values of the tuning parameters are recommended by Friedman, if the sample size is small). The estimated relative importance of each predictor variable in predicting the response based on the current MART model is shown in a plot [Friedman, 2000]. We selected the 3 variables with a relative importance of higher than .30 for the subsequent logistic regression analysis.
The number of patients used in the logistic regression analysis with these predictors was 151 (13% of the available patients were deleted), and 77,5% were correctly classified. The *AUC* was 0.801. In the 10-fold cross validation, the percentage correctly classified increased to 79,5%, and the *AUC* diminished to 0.767.

Categorical principal component analysis

CATPCA was applied to predictors of each domain separately to maximize the interdependence of predictors in one domain (for the domains, see Table 1), and in this way, to facilitate the selection process. In linear (logistic) regression analysis, we aim at a subset of predictors that are on the one hand not very highly correlated, and on the other hand most related to the response. For predictors in each domain (see Table 1), we performed a two dimensional CATPCA with both the predictors of the domain and the response. It is necessary to evaluate the predictor space in presence of the response, to enable the tracing of predictors that are strongly related to the response. To ensure a good representation of the response in a low-dimensional space, a larger weight (we chose 3) is attached to the response variable in the analysis. To apply the CATPCA

approach, continuous variables were made discrete by an optimal procedure based on [Max, 1960], implemented in SPSS version 10.0. To avoid dominance of a variable merely by its large number of categories, all variables are discretized into the same number of categories (we chose 7). CATPCA optimally scales the variables in the analysis, and we saved the transformed variables. Since a maximum of 11 predictors was allowed and we had 6 domains of predictors, we could select about 2 predictors per domain. We decided to choose predictors with a high correlation (› 0.50) with the dimension of the response (provided that the correlation between the predictors was not higher than 0.70; in such a case we selected the predictor with the least missing values). If such predictors were not available in a domain, we selected one predictor with a total variance accounted for (VAF) of 0.50 or higher (i.e. a predictor that dominated the solution), and the highest correlation with the dimension of the response.

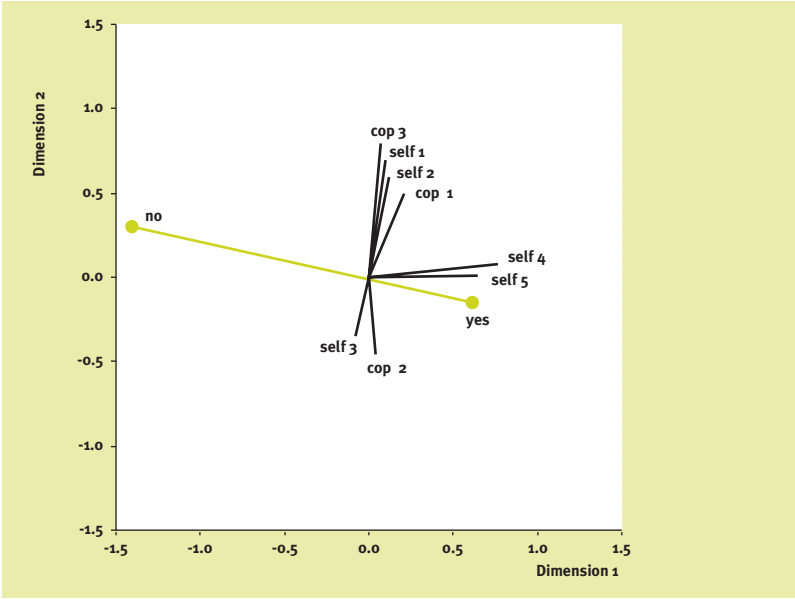


Figure 1
Two-dimensional CATPCA solution of the variables in the psychological domain. The arrows represent the correlations (i.e. the importance) of each predictor variable with the two dimensions. (self = self-efficacy; cop = coping); the dots represent the two categories of the response variable (no = continued smoking; yes = quit smoking). The dimension of the response is indicated with a dashed line.

On the basis of the two-dimensional CATPCA solutions, we chose seven predictors. In the psychological domain two predictors correlated highly with the dimension of the response (self-efficacy4 and self-efficacy5; see Figure 1), in the rest of the domains no single variable correlated highly. In these domains, we chose one predictor (having the highest correlation with the response dimension, and a total VAF of 0.50 or higher). The result of a logistic regression analysis with the seven selected predictors after optimal scaling is displayed in Table 3. The number of patients used in the analysis was 150 (13% of the available patients were deleted), and 79,3% were correctly classified. The *AUC* was 0.854. The percentage correctly classified decreased to 75,3% in the 10-fold cross validation, and the *AUC* to 0.802.

Results

Of the 173 patients who smoked before the cardiac event, 119 (68,8%) had quit smoking after the cardiac rehabilitation (M2; about three months after the cardiac event).

In all four solutions, two predictor variables were most important: specific self-efficacy and need of guidance to quit smoking (self-efficacy4 and self-efficacy5). From the results of the analyses, we may generate the following conclusions, which could be used for future research:

- patients who perceive themselves as less able to quit smoking at the pre-test have a higher risk of continuing to smoke after cardiac rehabilitation;
- patients who perceive themselves in need of guidance to quit smoking at the pre-test, have a higher risk of continuing to smoke after cardiac rehabilitation; and
- the three cardiac rehabilitation programs do not differ significantly in their stimulating effect on quitting smoking.

Table 3
Results of logistic regression analysis of quitting smoking on the seven selected predictors, optimally quantified by CATPCA (N = 150).
Notes: CI = Confidence interval.

predictors	odds ratio [4]	99% CI
work	1,25	0.67, 2.32
cardiac event	1,24	0.67, 2.30
self-efficacy4	3,35	1.52, 7.39
self-efficacy5	1,66	0.91, 3.00
quantity1	0,75	0.39, 1.44
quality of life (social)	0,66	0.35, 1.24
cardiac rehabilitation	1,62	0.91, 2.90

Our data mining results suggest that both CATPCA and MART are promising alternatives for selection of predictors to be used in a logistic regression analysis. The estimated logistic regression model using MART as preprocessing method has the highest cross-validated percentage correctly classified (79,5%), which is 10% higher than the prior class probability. The estimated logistic regression model using CATPCA as preprocessing method has the highest cross-validated *AUC* (0.802), which indicates a moderately good prediction. Also, fewer patients are deleted from the analysis with MART and CATPCA than with the stepwise selection procedure and decision tree analysis. Furthermore, in the CATPCA procedure, we combined expert knowledge with the use of a data mining technique, by applying CATPCA in each domain of predictor variables. If such expert knowledge had not been available, CATPCA could have been used to determine separate domains as well.

4 The odds of quitting smoking equals the number of patients who have quit smoking divided by the number of patients who have not, which is 103/47 = 2,19. The odds ratio is interpreted as follows: the odds of quitting smoking increase multiplicatively by the odds ratio for every unit increase in the predictor. The effect of the predictor is not significant if the confidence interval includes 1,00.

Discussion and challenges for the future

The application above shows different prediction performances of four data mining methods in the analysis of a real data example. One limitation of our results is, obviously, that some of the differences can be attributed to the specific data example used. In other words, we could have found somewhat different results if we had used different data. Another limitation of the results is that the conclusions about the performance of the four methods are drawn on different populations (the final sample sizes used in the logistic regression analyses varied between 108 and 151). The comparison would still be unbiased, if we may assume that the missing values occurred at random. However, some missing values in the present data were systematic. For example, the variable quality3 (satisfaction with family support) had a lot of missing values, because the item was not applicable to patients without a family. One reason for the existence of different sample sizes in the analysis is that the methods differ in the way they handle missing data. Backward stepwise regression and decision tree analysis (the implementation in Splus) delete all cases with one or more missing values on any predictor variable, whereas MART and CATPCA use all available cases. A second reason is that different predictors (with more or less missing values) are selected by the different methods. The use of an imputation method (e.g. the EM algorithm [Dempster, 1977]) before the selection of variables would be a possible solution to the missing data problem.

The data example used in this paragraph is very small compared to the data sets commonly used for data mining, which may contain over 70 billion observations [Hand, 2000]. The aims of this study, however, were to illustrate the application of data mining tools to behavioral science data that typically contain few subjects (less than 1,000) and a large number of variables, and to relate the various steps of the knowledge discovery in databases process to those of the empirical scientific process. The exploratory nature of the analyses, that is, the search for interesting information without a priori hypotheses, was essential to this study.

Each of the four methods we used has its own strength. Challenges for the future are on the one hand to integrate the strengths of two or more methods into one single method, and on the other hand to get more knowledge about which method is more appropriate for which problem. Of course, different methods can, and should be, applied successively to the same data to compare the results.

At the moment, we are investigating the integration of a tree-based method (regression trees) into linear or categorical regression analysis [Dusseldorp, 2001]. The strength of a tree-based method, that is, the automatic detection of interactions between predictor variables, is combined with the linear regression method, to estimate the significance of the interaction effect(s) traced by the

tree-based method. The resulting method is called the Regression Trunk Approach, and is especially appropriate for studies with many predictor variables, having no a priori hypotheses about the type of interactions that are contained in the data.

References

- American Psychological Association. (1994). Publication Manual of the American Psychological Association (4th ed.), Washington, DC
- Breiman, L., J. Friedman, R.A. Olshen, C.J. Stone. (1984). Classification and Regression Trees. Wadsworth, Belmont, CA
- Cattell, J. (1890). Mental Tests and Measurements. Mind **15**:373-81
- Clementine. (1998). Clementine User Guide. Version 5. Integral Solutions, Basingstoke
- Cronbach, L.J. (1951). Coefficient Alpha and the Internal Structure of Tests. Psychometrika **16**:297-334
- Dempster, A.P., N.M. Laird, D.B. Rubin. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Ser. B **39**:1-38
- Dusseldorp, E., J.J. Meulman. (2001). Prediction in Medicine by Integration of Regression Trees into Regression Analysis with Optimal Scaling. Methods of Information in Medicine **40**:403-409
- Elderen, T. van, M. Chatrou, H. Weeda, S. Maes. (1997). Leidse Screening Vragenlijst voor Hartpatiënten (LSV-H; Leiden Screening Questionnaire for Cardiac Patients). Clinical and Health Psychology Section, Leiden University
- Elderen, T. van, E. Dusseldorp. (2001, in press). Lifestyle Effects of Group Health Education for Patients with Coronary Heart Disease. Psychology and Health
- Freund, Y., R.E. Schapire. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences **55**:119-139
- Friedman, J.H. (1999). Greedy Function Approximation: A Gradient Boosting Machine [Submitted]. Available: http://www-stat.stanford.edu/~jhf.
- Friedman, J.H. (2000). Tutorial: Getting started with MART in Splus [On-line paper]. Available: http://www-stat.stanford.edu/~jhf
- Galton, F. (1869). Hereditary Genius. Macmillan, London
- Galton, F. (1886). Regression towards Mediocrity in Hereditary Stature. Journal of the Anthropological Institute of Great Britain and Ireland **15**:246-263
- Gaul, W., F. Säuberlich. (1999). Classification and Positioning of Data Mining Tools. In: W. Gaul, H. Locarek-Junge. (eds.). Classification in the Information Age. Studies in Classification, Data Analysis, and Knowledge Organization. Springer Verlag. pp145-154

- Hand, D.J., G. Blunt, M.G. Kelly, N.M. Adams. (2000). Data mining for Fun and Profit. Statistical Science **15** (2):111-131
- Hand, D.J., R.J. Till. (2001). A Simple Generalisation of the Area under the ROC Curve for Multiple Class Classification Problems. MachineLearning, **45**:171-186
- Hillers, T.K., G.H. Guyatt, N. Oldridge, J. Crowe, A. Willan, L. Griffith, D. Feeny. (1994). Quality of Life after Myocardial Infarction. Journal of Clinical Epidemiology **47** (11):1287-1296
- Max, J. (1960). Quantizing for Minimum Distortion. IRE Transactions on Information Theory **6**:7-12
- Menard, S. (1995). Applied Logistic Regression Analysis. Sage, Thousands Oaks
- Meulman, J.J., W.J. Heiser, SPSS Inc. (1999). SPSS Categories 10.0. SPSS Inc., Chicago
- Stevens, J. (1992). Applied Multivariate Statistics for the Social Sciences. Second edition. Lawrence Erlbaum, Hillsdale
- Thorndike, E.L. (1913). Educational Psychology. Vol. 2. The Psychology of Learning. Teachers College, Columbia University, New York

2

2.4

Conclusions

Jeroen Meij

General

Looking at the next decades, we have to extrapolate the developments sketched in the previous chapters. User-centered systems, standardization and access to data and computing power are recurring needs in many fields of science. Like the distributed nature of data on the Internet and through connected databases, computing (analytic) power will be increasingly distributed in what is called a Grid. This will eventually mean that scientists can rent or borrow the necessary computing power and software when they need it from the Grid.
As Grid based systems advance, their analytical tools will become modular, enabling individual scientist's approaches to be implemented easily.
Through the general increase in computing power we will see an increase in real time analytical processing, performing several steps even between the collection and the storage of data.
Analysis of spatial, temporal and relational data is generally considered an important area of development.
Data mining technology will play a large role in the modeling of natural language, ultimately leading to automatic extraction of meaning, automatic summarization and classification of text.
Learning with data mining techniques and agent technology will play an important role in many areas.

Domain knowledge

Availability of scientific documents

Many publications are being made available on-line, by the authors themselves and by publishers (paid, with password access) or other organizations. Regrettably, a lot of these publications are only accessible with paid memberships. For most western universities this is no problem, but for many developing countries it is. Another unwelcome development is the publication of documents in print-directed formats as Postscript and Portable Document Format (PDF). These can be downloaded and viewed, but are very hard to analyze on-line[1], because of the lack of distinction between content and shape. Most search engines do not take the contents of these documents into account.
To advance the availability of on-line documents, a standardized searchable format should be adopted. XML (eXtensible Markup Language) is a contender for the preferred future format, since it incorporates a strict division between content and shape. Moreover, especially in the case of books, some form of copyright management has to be agreed upon. Since production and distribution costs for digital documents are negligible, digital documents should be substantially cheaper than printed books.

Search

Intelligent search methods will aid scientists in their search for knowledge, meta-search and agent technology being the first steps. Mapping technology will visualize scientific fields and the connections between them. Also the dynamics of scientific fields can be visualized, perhaps leading to the discovery of new themes.
Interactive visualization techniques for search results will reveal relations between scientific publications and terms, and help understanding related domains (see also Chapter 5.6, Web mining).
Agents will become more and more autonomous, representing the scientific interests of the owner, but also anticipating his needs, learning through experience. Since the Internet is constantly changing, we would expect scientists to build and manage their own digital libraries, either on-line or off-line, privately or with colleagues.

Scientific methodology

The common hypothesis-testing procedure of science is and will be increasingly complemented with hypothesis generation, especially in areas with large quantities of data (See Chapter 2.1, Figure 2).
In many areas, data and interactions between variables are so numerous and complex, that other methods are infeasible. Needless to say, a division of the available data between a training set and a test set is a prerequisite, as is thor-

1 Although progress is made in this direction.

ough hypothesis testing, followed by confrontation with domain knowledge. Only from a constructive sequence like this, understanding can arise.

Expectations for science areas

Life sciences

The choice of this title is a direct consequence of the converging areas of biology and bioinformatics, medical science and medical informatics.
Starting with bioinformatics, we will see a significant increase in the available data on the expression of genes. Also a further understanding of the interactions between genes, proteins and metabolites will be achieved. It will come as no surprise that bioinformatics will play an increasing role in the development and application of drugs. The integration of chemical, biological and clinical data will be a key factor for development. On the application side pharmacogenomics will increase in importance by tailoring drugs for patients with a specific genetic profile.
Data mining for drug discovery will be evolving towards on-line processing with some decision making and acting capabilities, for instance when the analysis needs additional tests to evaluate competing hypotheses (that have just been discovered). On the other hand an integration of numerical and symbolic data mining will enable us to compare and trade off between (predictive) accuracy and understandability. Ideally, as in many other areas, research will be performed by multidisciplinary teams of data mining and domain experts.
Closely related to this, the implementation of electronic patient records and hospital information systems is absolutely necessary for further advancement of knowledge discovery from medical data. The data should not be limited to numerical data, but should also contain results from physical examinations and images or 3D scans. For a rapid advancement of medical knowledge the standardization of hospital data (i.e. by ISO TC215) is imperative, especially in the light of bioinformatics. Hypothesis discovery, temporal diagnostic-pattern discovery, short and long-term therapeutic effects assessment and discovery of new diseases are all expected to follow from medical knowledge discovery in the next decades. Analysis of temporal and relational data can be considered especially important in this field, and both areas will require a lot of development.
Another important issue is the integration of domain knowledge and data derived knowledge. Successful integration will lead to reliable decision support systems that can be used for education and to assist with diagnosis for experts and patients. In the latter case patients could use web based systems to decide whether they should consult a doctor, and which hospital is best suited.

Environmental and ecological sciences

Biodiversity research

All over the world, museum collections of organisms are being entered into databases, largely uncoordinated. GBIF, the Global Biodiversity Information Facility of the OECD could provide a remedy for this situation.
Societal applications range from biodiversity conservation to agriculture, fisheries, health, forensics, environmental management, and so forth.
Developments from bioinformatics will join with the biodiversity databases to enable effective knowledge discovery. Eventually, these databases will also incorporate image (and 3D) information. Mining tools should be user oriented, fast and text- and image capable.

Environmental

For the environmental sciences, a need can be identified for data access and preparation, supported by distributed computing power. The demands of this area are, thus, closely related to the meteorological demands mentioned in Section 2.2.5. Special attention should be given to spatial and time-series analysis, from data representation to knowledge representation. A strong need for standards for spatio-temporal models can be observed, which are expected to lead to implementations in current GIS and database systems. Again, user-centered analysis systems are required to visualize and model these concepts. Data enrichment should become a standard database feature, for instance with the outcome of rules derived from the database. Finally, from all this, agents could help in investigating the interaction of different models, thus assisting in the decision making process.

Ecosystem management

In a somewhat smaller scale of ecosystems with relatively well defined boundaries, predictive models and visualization techniques are already being employed. River management is a living example. Many other areas are expected to deploy the same techniques.

Economics

Growing quantities of microdata, for instance about the buying behavior of consumers will provide new opportunities for data mining in economics. Given that training and test sets are available with sufficient data, data mining can be a useful tool for hypotheses generation in economics. If we add the development of agents to this, we can train agents based on historical data. With these agents we can simulate social processes through emergent behavior, thus creating adaptive social simulation systems, useful for practical and fundamental

research purposes. We could derive market behavior, observe social trends and market mechanisms and simulate the consequences of political measures.

Closing remarks

In this part of the book we have seen many examples of knowledge discovery from data in science. Summarizing in a few keywords, we expect to see:

- integration of functionality and fusion of databases
- agents assisting in acquiring domain knowledge
- integration of data-derived knowledge and domain knowledge
- distribution of data and computing power.

However, some obstacles will require our attention:

- restricted access to domain knowledge (standards and intellectual property)
- lack of standardization of data formats within domains
- poor user-friendliness of systems.

As the trends mentioned above become reality – provided these obstacles are negotiated – science will be advancing faster then ever before.

3

Knowledge Discovery in Business and Government

3.1 Introduction

This part aims to build a bridge between the world of the manager, policy maker or marketeer and the world of the data miner. For this purpose we have tried to formulate business actions (needs) that can be performed – for a large part – by data mining.
The results of these actions can provide useful support for making business decisions. Technically speaking there may be several different data mining actions which can be performed to achieve the end goal of the business action. Prediction for instance may involve segmenting, classification and modeling as steps in the process.
In this part we will provide cases — real and fictitious — of situations where data mining is or could be used to reach business or government goals.

The data mining tasks mentioned in this part are covered in more detail in Part 6, so we will only give short descriptions in the paragraphs, with links to Part 6 wherever appropriate.

Business actions

Although the terms used below sometimes coincide with names of data mining methods, we specifically refer to the 'real world' actions (or needs) here. We have formulated example questions a manager might ask himself to illustrate the business actions.

Grouping (clustering)

What distinct groups exist within my customer base?
Being able to discern a group of entities with common characteristics. From a mass of data one ore more useful groups are identified (see Section 6.2.6). Examples might be customer groups or demographic regions.

Categorization (classification)

Which group does this new customer belong to?
Assigning an entity to a known category (see Section 6.2.7).
Examples might be assigning customers to a known group, separating different quality classes of fruit, separating different vehicle types.

Detecting

If I only had a warning light, indicating we should investigate these particular cases...
Being able to detect a deviating state that can be considered relevant. Intrusion detection in a network, phone or money-transfer patterns that indicate fraud can be seen as examples. Detecting may be regarded as a special form of classification.

Modeling

What are the influences of family size, income and...? on choosing a car?
Generating an abstract description of (a part of) reality.
This mainly concerns the cases where we are interested primarily in understanding processes. With this understanding we can develop better regulating mechanisms or products.

Prediction

What car models will this man be interested in?
Predicting behavior of groups, individuals or systems.
When we have a model, we can also predict the outcome for new values, provided the process itself does not change. We can predict which clients will be interested in a caravan policy, or will be paying their credit card debts in time.

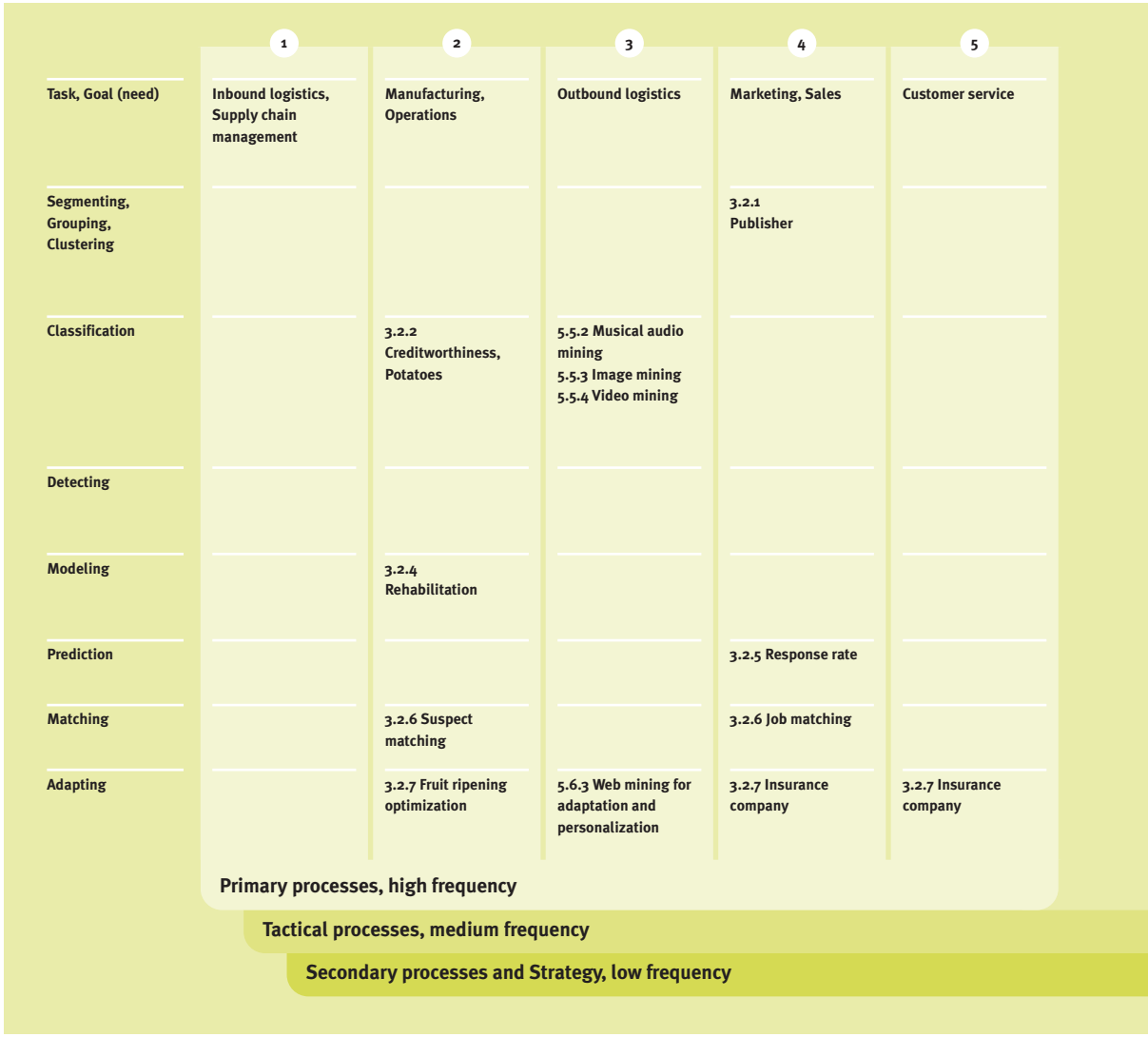


Figure 1
Business tasks and cases. Please note that an empty cell does not mean that a combination is not possible. Whenever a case describes multiple business tasks, it is placed in the lowest row, since the tasks in the upper rows often precede the tasks in the rows below them.

Matching

Which offers are able to fulfill this request?

Creating a useful link between two or more entities. Examples might be job-matching, purchase/sales matching or dating.

Adapting

What should our homepage look like, when we are visited by a sports enthousiast? and when a teenager visits the page?

Being able to adapt a system to a situation (or customer). Examples might be web pages, educational systems and procedures.

6	7	8	9	10	
Monitoring, Tactics, Control	Support purchasing	R&D, Technical infrastructure development	Knowledge management Human resources management	Corporate strategy, Structure	Task, Goal (need)
			2.2.1 Text mining 5.4 Text mining 5.7 Personal knowledge management		Segmenting, Grouping, Clustering
5.5 Multimedia mining		2.2.1, 2.2.2, 2.2.3, 2.2.4 General application for science 5.5.3 Image data mining tasks	2.2.1 Text mining 2.2.2 Agents 2.3.1 Medical science 2.3.2 Medical diagnosis	2.2.3 Science mapping	Classification
3.2.3 Suspect behavior, Waste transport					Detecting
2.3.9 River management		3.2.4 Burglary 2.3.8 Environmental sciences		2.3.6 General economics	Modeling
3.2.5 Financial markets				2.3.7 Microeconomics, emergence	Prediction
			3.2.6 Job matching		Matching
			5.7.2 Personal knowledge management	2.3.8 Environmental sciences	Adapting

The matrix given in Figure 1 indicates the cases discussed in Parts 2, 3 and 5, their business actions, and the corporate units involved. As shown here, the applications of data mining spread through most business columns.

In Section 6.1.1 we offer a more technical view of the needs/tasks discussed here. The diagram given in Section 6.1.1 links the needs to data mining techniques discussed in Part 6.

3

3.2 Cases

3.2.1 Segmentation, clustering

Introduction

Segmentation is often defined as 'being able to discern a group of entities with common characteristics'. From a mass of data one or more useful groups are identified. Examples might be customer groups or demographic regions. Segmentation is closely related to classification and is very useful for marketing purposes. More on the technical details of segmentation and classification can be found in Section 6.2.6, Clustering.

Advertising strategy discovery: publisher

Peter van der Putten[1]

In this section we describe a case in which a large Dutch publisher, De Telegraaf, uses customer profiling and segmentation to promote advertisement sales. In a marketing context, it is often tacitly assumed that customer databases are mined for knowledge. However, due to their rich nature (many fields) surveys are even more likely candidate data sources to be mined. The Telegraaf case is a very good example of this approach.

Business problem: promoting advertisement sales

De Telegraaf is one of the major publishers in the Netherlands. The Marketing Services Department is responsible for marketing of the largest Dutch daily newspaper 'De Telegraaf' and the magazines of the 'Telegraaf Tijdschriften Groep', which include various titles on music, lifestyle, sports, cars, etc. that are aimed at different target groups (like Autovisie, Elegance, Privé, Man, Hitkrant, Oor). A major problem for a publisher is that they only get contacted at the end of the advertisement chain, when the decision on where to advertise has already been made. Typically, for a certain campaign an advertising agency will design a media strategy and an appropriate communication message and hire a media agency to work out a detailed media plan and buy advertising space. By supplying large potential advertisers with high quality and serendipitous information about their target groups, De Telegraaf can get involved and grab some attention in the very beginning of the advertisement chain, maybe even before an advertising agency has been contacted. Furthermore, these target group reports are a win-win loyalty marketing tool to strengthen the relationship with major advertisers.

The target group reports are based on the so-called SUMMO Dutch National Media Survey, a rich collection of data on lifestyle, product and media consumption of Dutch consumers, which is carried out on an annual basis. In all major countries around the world similar national media surveys are available. For the SUMMO survey over 13,000 consumers are interviewed. The surveys contain over 5,000 answer codes (possible answers). Since 1994 De Telegraaf uses data mining technology to quickly discover relevant, interesting and surprising knowledge about target groups (see also [Putten, 1999]).

The standard statistical methods that De Telegraaf used before (cross tables) only offered the opportunity to calculate the frequencies of answers for a certain specific question, for example the frequency of Nike customers versus age or versus choice of newspaper. Given the enormous number of questions and

1 Drs P. van der Putten, pvdputten@hotmail.com. pvdputten@liacs.nl, Leiden Institute of Advanced Computer Science, Leiden University, Leiden, The Netherlands

answers in the survey, this approach leads to practical problems. There is no time to screen all relevant questions and answers in the survey. Furthermore, the marketeers of De Telegraaf lack a priori knowledge about the target group. So even if unlimited time was available, even the most experienced marketeer might miss some important profiling characteristics of a target group. In practice, marketeers only have very limited time to prepare and present a report (one to two days!) and the goal is to surprise the potential advertiser with new knowledge, which is quite ambitious, if the client has years of experience with the target group. So it was desirable that instead of asking a marketeer to generate and test all interesting hypotheses, a mining tool should be used to generate potentially interesting profiling characteristics and segments[2].

Data mining solution: profiling and segmentation

Data mining profiling analyses offer a solution to the problem. First a target group needs to be selected by use of standard strict criteria or fuzzy matching criteria, e.g. 'Sneaker=Nike' or 'Strict: Female, Fuzzy: around 50, highly educated, likes gardening & cooking, etc.'. Then the profiling engine uses standard statistical tests to generate a profile of deviating characteristics, by computing all frequency distributions of answers in the target group and comparing these frequencies with a reference group, for example the whole Dutch population. As an example the top profiling characteristics of vodka drinkers are displayed in Figure 1.

Profile Analysis

TotalProfile | Variable | Marks — 'Drinks more than once a month (alc) Wodka' vs. 'Entire Population'

Mask: <All variables> Score>=10 count>=10

	Variable^	Feature	Scr-	N (146)	#Drinks	%Drinks	#Entire	%Entire
1	Drinks more than once a month (alc)	Wodka	97.3	146	148,466	100.0%	148,466	2.1%
2	Grants for people 13/17 years^	Yes, respondent receives study allowance	55.8	14	18,728	77.5%	335,715	21.8%
3	Drinks more than once a month (alc)	Whiskey	39.5	74	70,072	47.2%	557,816	7.7%
4	Visits > 1 a month	liquor store	37.2	85	87,480	58.9%	1,573,965	21.7%
5	Drinks more than once a month (alc)	rum (brown/white)	36.5	55	61,490	41.4%	329,048	4.5%
6	Benefits^	Scholarship/Grant	32.9	14	18,728	42.5%	335,715	9.6%
7	Visits > 1 a month	cd/record shop	32.8	84	90,628	61.0%	2,047,243	28.3%
8	Sex^	man	31.6	110	117,760	79.3%	3,454,061	47.7%
9	Watching kinds of TV programs	science fiction films	28.4	92	99,110	66.8%	2,776,675	38.3%
10	Does activity > 1 a month	calling information numbers/0900 numbers	27.5	67	67,158	45.2%	1,286,274	17.8%
11	Primary people watching kinds of TV progs	mainly action films	26.5	85	92,284	62.2%	2,585,161	35.7%
12	Drinks more than once a month (alc)	Liquors	26.4	47	54,111	36.4%	724,617	10.0%
13	Consumes salty snacks	other crisps (nachos etc.)	26.3	79	82,951	55.9%	2,140,899	29.6%
14	Buy clothes which are on offer	enjoys looking in shops for audiovisual eqpt	24.9	64	76,726	51.7%	1,940,106	26.8%
15	Pays attention to	cinema advertising	24.8	72	76,119	51.3%	1,919,835	26.5%
16	Self monitoring^	high	24.5	78	80,268	54.1%	2,144,024	29.6%
17	Personal possession various items	walkman	24.4	82	89,383	60.2%	2,589,200	35.8%
18	Current additional income^	Yes	24.4	18	21,830	39.9%	582,102	15.5%
19	Visits > 1 a month	shop at gas station	24.2	44	54,301	36.6%	892,628	12.3%
20	Personally interested	economy	23.6	86	88,562	59.7%	2,610,389	36.1%
21	Opinion asked regularly	audio appliances	23.4	46	53,422	36.0%	907,881	12.5%
22	Watching kinds of TV programs	horror films	23.4	77	88,069	59.3%	2,599,922	35.9%
23	Interest in various subjects	movies	23.4	83	84,363	56.8%	2,423,485	33.5%
24	Reason not employed^	No. studies	23.0	19	24,173	44.1%	786,780	20.9%

Figure 1
Target group profile for vodka drinkers. For example, watching science fiction movies is one of the most profiling characteristics, because 66,8% of vodka drinkers likes these movies, compared to 38,4% of the Dutch population, resulting in a selectivity index of 28,4 (numbers slightly changed for copyright reasons).

The target group profile offers a description of the average vodka drinker. However, this is only part of the solution. It may well be that the average vodka drinker does not exist, but that there are at least a small number of different vodka drinker types. The borders between these segments are typically not strict. Each of these customer types needs to be addressed with a different communication strategy (in other words: there are always opportunities for De

2 The solution was based on the DataDetective Visual Datamining Environment, developed and marketed by Sentient Machine Research.

Telegraaf). Such a differentiated strategy results in both a higher efficiency and effectiveness. By using interactive data mining segmentation tools, the marketeer can discover which segments exist within the target group.

The segmentation process is carried out in a number of steps. First the target group, say all vodka drinkers, are spread out randomly on a flat plane. During the step-by-step segmentation process each vodka drinker moves a small distance towards those neighbors in the plane whose characteristics match best.

Figure 2
Vodka drinker segmentation. Three segments can be distinguished, each with a distinct profile of characteristics (photos for illustration purposes only).

The end result is a plane with several segments (Figure 2): within a segment customers are as similar as possible and the segments are as different as possible. The match is not determined on just two variables such as gender and age, but on all questions and answers from the survey, or any subset of these questions, if desired.

The profiles of the various segments tell a different story from the profile of the average vodka drinker. Three segments can clearly be distinguished.

- A group of teens and adolescents, both boys and girls, who drink vodka with orange juice and cola, like to go out, watch videos and are interested in house music. This group does not read too much, apart from De Telegraaf and teen music magazines, but they watch advertisements in the cinema and in the street and they like to watch TV.
- A group of twenty-something year olds, students and young couples, who live in rented apartments and buy products like baby shampoo, fresh vegetables and cat food. They read more liberal newspapers like the Volkskrant and opinion magazines.
- A group of older business men; they own credit cards and other financial products, they drive around in lease cars more often and read conservative newspapers and magazines. They don't watch TV that much.

It is obvious that the profile of the average vodka drinker only presents a limited view. The segmentation exercise offers a richer view of the target group, which can contribute to better brand positioning and a distributed media strategy.

Vision for the future

For the past seven years De Telegraaf has been using descriptive data mining, and more specifically profiling and segmentation, as a major instrument for offering services to potential and actual advertisers, resulting in better advertisement sales. The success of this formula has been the integration of data mining in a core business process. De Telegraaf has a clear vision of how the data mining results should be exploited. The impact of the technical implementation has been minimized by updating the central survey database on a yearly basis.

Of course this customer profiling and segmentation approach can be applied (and has been applied) for a multitude of other purposes, not limited to surveys, such as insurance risk analysis, policy research, market basket analysis, customer relationship management, crime analysis and medical discovery.

From an application point of view it might be interesting to spend some future technical research on extending the rather straightforward univariate profiles to multivariate profiles that are generated by association rules and decision trees or non prepositional profiles that can be constructed by inductive logic programming algorithms. However, although these profiles are more expressive and complex, users might still prefer simple profiles! For segmentation it might be interesting to do some technical research on more immersive visualization technology such as three-dimensional segmentation and visualization. Again, the simple two-dimensional solutions will be difficult to defeat in terms of clarity.

Finally, for descriptive data mining tasks such as profiling and segmentation, the richer the data, the better it is. This is in contrast to prediction tasks that suffer more from the curse of dimensionality. However, rich data is expensive, and, generally, rich data is known only for subsets of customers. A new hot field in data mining concerns data fusion methods that allow us to enrich entire customer databases with survey information that is only available for a sample, in other words, carrying out a virtual survey with each customer [Putten, 2000]. If this technology becomes mature, a whole new arena for segmentation will evolve.

References

- Putten, P. van der. (1999). Datamining in Bedrijf. Informatie en Informatiebeleid **17**:3
- Putten, P. van der. (2000). Data Fusion: A Way to Provide More Data to Mine. Proceedings 12th Belgian-Dutch Artificial Intelligence. Conference BNAIC'2000, November. De Efteling, Kaatsheuvel, The Netherlands

3.2.2 Classification

Classification or categorization can be defined as 'assigning an entity to a known category'. Examples might be assigning customers to a known group, separating different quality classes of fruit on the conveyor belt, separating different vehicle types as they drive by.
The first article concerns the assessment of creditworthiness of companies by visual classification, the second article describes the classification of potatoes from a 'product quality' point of view. Since quantities are very large, speed was an important factor here. More on theory and algorithms of classification in Section 6.2.7, Classification.

Visual assessment of creditworthiness of companies using Self-Organizing Maps

Roger P.G.H. Tan[1], *Jan van den Berg*[2], *Willem-Max van den Bergh*[3]

Self-Organizing Maps

This book discusses several possible data mining techniques. In this chapter we use the Self-Organizing Map (SOM, extensively treated in [Tan, 2000 and Kaski, 1997], included on the CD-rom) to visualize and cluster companies according to their financial statement. We aim to characterize these clusters of companies in terms of creditworthiness.

Other examples of the use of SOM can be found in very different areas of science and business processes. For example, a Self-Organizing Map can be used to group countries in the world according to macroeconomic figures. Interesting differences between emerging and developed countries are clearly visualized on the SOM display [Deboeck, 1998]. But it is also possible to use the SOM on the medical domain, grouping patients according to specific characteristics [Tan, 2000, Appendix III).

Many different software implementations of the SOM are available. We have used the Viscovery SOMine application, built by Eudaptics in Austria [Eudaptics, 1999].

Credit rating classification using Self-Organizing Maps

Introduction — problem domain

When banks or companies lend money to other companies or governments, it is often in the form of a bond, issued by the borrowing company. The buyers of the

1 Drs R.P.G.H. Tan, tan@mediaport.org, Robeco Group N.V., Quantitative Research Department, Rotterdam. This article is mostly based on Tan's masters thesis [Tan, 2000]

2 Dr Ir J. van den Berg, vandenbergh@few.eur.nl, Erasmus University Rotterdam, Faculty of Economics, Rotterdam, The Netherlands, http://www.eur.nl/few/people/jvandenberg/

3 Drs W.-M. van den Bergh, Erasmus University Rotterdam, Faculty of Economics, Rotterdam, The Netherlands

bond have to make an assessment of the creditworthiness of the issuer, based on the financial statement of the issuer (balance sheet and income account) and on expectations of future economic development.

Most buyers of bonds do not have the resources to perform this type of difficult and time-consuming research. Fortunately, so-called rating agencies exist which specialize in assessing the creditworthiness of a company, using a combination of a quantitative and a qualitative analysis. The resulting credit or bond rating is a measure of the risk of the company not being able to pay an interest or redemption payment of its issued bond. These ratings are often seen as representative for the market opinion of the creditworthiness of an issuer.

In this chapter we will try to group companies according to their financial characteristics as expressed in the financial statements of these companies. The found and visualized clusters can then be examined to find (perhaps previously unknown) patterns in the data. This leads to a characterization of the creditworthiness of the companies in each cluster. The found clusters can then be compared with the ratings as perceived by the S&P rating agency, which in our case functions as the 'market' opinion on creditworthiness. A third step involves using the found knowledge to construct a credit rating classification model, aimed at correctly classifying the companies according to S&P rating.

Used data

Financial figures are generally quite volatile and often contain a lot of errors. Financial statement data in particular does not always reflect the true state of a company, because companies try to look their best by using a multitude of accounting practices (this is known as 'window dressing'). Several data sources provide company level financial figures, and most data sources fortunately ensure that all figures are calculated using similar accounting practices. So by limiting the data to a single source we at least know that the figures are mutually comparable with respect to the accounting basis. However, care should still be taken to preprocess the data correctly (check for inconsistencies and remove outliers) before attempting to extract knowledge from the data.

The used data per company consists of a selection of 18 different financial ratios. Each ratio provides a summary of a specific aspect of the financial statement of the company. The 300 companies in our universe all reside in one sector, so the financial ratios for these companies are mutually comparable (a bank will have a very different financial statement from a construction company). We have examined one cross section of quarterly data in 1998.

Main research questions

The following questions will be answered in this chapter:

- Is it possible to make an assessment of creditworthiness of companies using Self-Organizing Maps, based on the financial ratios for these companies?
- Does this assessment of creditworthiness coincide with the creditworthiness as perceived by the S&P rating agency?
- Can we use the found knowledge to create a credit rating classification model?

Visualizing the financial companies landscape

Creating and evaluating maps

Using the SOM technique the data mining step of the knowledge discovery process [Fayyad, 1996] boils down to creating and evaluating maps.

First the observations (in our case companies) are projected onto the Self-Organizing Map, using a process called 'self-organization' (see [Tan, 2000, chapter 3]). Then the projected observations can be clustered by means of a bottom-up clustering algorithm. It is now possible to evaluate the clustering found using available information like summary statistics per cluster, the distribution of each individual variable over the map (in submaps called 'component planes') and specific financial domain knowledge.

This process is often performed iteratively. After creating a map, the new insights provided by interpreting the map (comparing the main map and the component planes) and evaluating the results may force us to create another map using different settings or variables. In the end, the clustering found

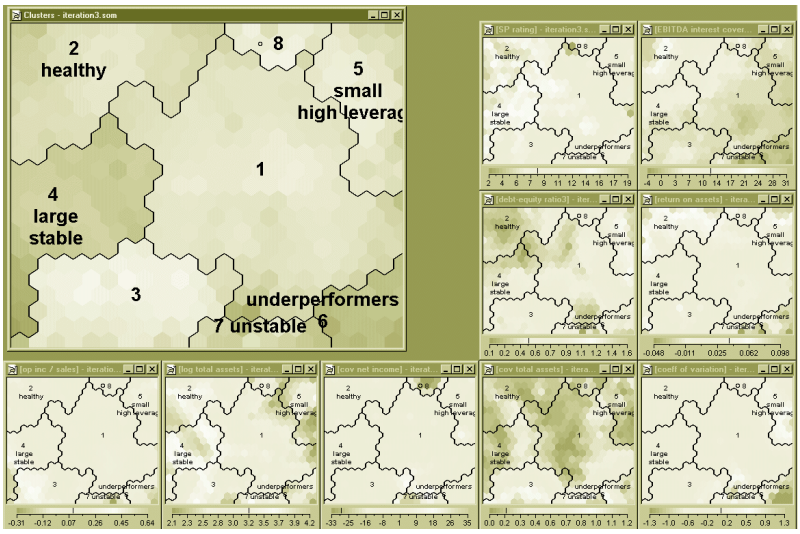


Figure 1
Company clustering and component planes.

should accurately reflect the key (not necessarily linear) relationships between clusters and individual variables.
The full SOM display, showing the main map and the component planes for each variable, is shown in Figure 1. Based on this display the following characterization of the clusters can be made[4]:

C2:	Healthy companies with high interest coverage, low leverage, high profitability, very stable companies and low perceived market risk. Remarkable: these are not always the biggest companies.
C4:	Large stable companies with a high profit margin. Remarkable: not so high interest coverage.
C1, C3 and C8:	Average companies with no real outstanding features.
C5:	Small companies with low interest coverage and high leverage. Remarkable: a stable coefficient of variation of net income, these companies do not grow much.
C6:	Underperformers: very low interest coverage, very low or even negative profitability, negative earnings forecasts.
C7:	Unstable companies: very unstable and a very high perceived market risk.

Using domain specific knowledge the companies in the 'healthy' and 'large' clusters would be designated as having high creditworthiness, while the 'underperformers' and 'unstable' companies would be designated as having low creditworthiness. The companies in the 'average' clusters are expected to have average creditworthiness.

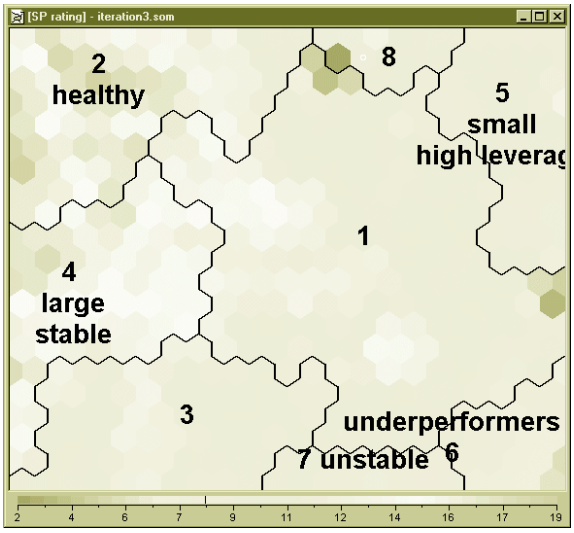
Matching model creditworthiness and market creditworthiness

The found relationship between financial ratios and creditworthiness can be visually verified using the S&P rating for each company. The rating is shown as a specific component plane, with red colors designating a higher rating and blue colors designating a lower rating. An overlay of the found clusters (Figure 2) shows the match between the creditworthiness as found by our model and the creditworthiness as perceived by 'the market' (the S&P ratings). Please note that while our clustering is based on financial ratios alone, the S&P rating is based on a quantitative and a qualitative analysis [S&P, 2000].

Visual inspection of Figure 2 reveals that the assessment of creditworthiness based on financial ratios alone forms a good approximation to the creditworthiness as perceived by the market, expressed in the rating. Clusters earlier designated as having a high creditworthiness ('healthy' and 'large, stable') mostly contain companies with a high credit rating. Clusters earlier designated as having a low creditworthiness ('small', 'unstable' and 'underperformers') mostly contain companies with a low credit rating.

4 Please refer to [Tan, 2000] for a full description of the used variables.

Figure 2
Matching clusters and ratings [Tan, 2000].



It is possible to define a more quantitative measure for the goodness of fit[5] of the ratings mapping. The specific definition lies outside the scope of this article (more information can be found in [Tan, 2000, chapter 4]), but suffice it to say that approximately 60% of the ratings is correctly mapped on the found clustering. So when we assume the following:

- The used financial ratios are a representative financial characterization of the companies in this sector.
- The SOM creates a good model of the underlying data.
- The data does not contain any major errors.

then approximately 60% of the rating of a company can be explained by its financial statement. It is difficult to attribute the other 40% to specific factors, one possible explanation could lie in the qualitative analysis performed by the rating agency that is not included in our model.

Creating a (credit) rating classification model

Model construction

The found knowledge can be utilized to construct a credit rating classification model, using the SOM as a non-linear, semi-parameterized regression model [Bishop, 1995]. The data used is similar to the previously used data, but this time we split the data into a train and a validation set. In addition, we apply the technique of semi-supervised learning, as described in [Eudaptics, 1999] and [Tan, 2000, chapter 3].

After training using the train set, the different areas of the SOM function as proxies for companies in different financial situations. Their associated ratings convey the common credit outlook S&P employs for these kinds of companies.

5 Measure of how well a rule or hypothesis fits a set of observations.

When we want to evaluate this common rating for 'new' companies in the validation set, we should perform the following steps:

- Evaluate the specific financial situation for this new company (and thus determine the placement of the company on the map).
- Read the associated common S&P rating off of the map.

Clearly, we should not expect this kind of model to be able to exactly assign the correct S&P rating to a company. The S&P ratings are based on a quantitative and a qualitative analysis, whereas our model assigns ratings based only on quantitative information.

Model validation

We can calculate different measures for the performance of the classification model on the validation set. One of them is the R^2 of the supposed linear model after classification of a number of new companies. A comparison between different models (linear and non-linear) is now possible. The achieved R^2 is 0,65, which can be interpreted as an adequate performance.

The relative performance of the model can be evaluated by comparing the performances of different models. A full comparison between SOM, linear regression and ordered logic [Tan, 2000, chapter 5] shows that this time no substantial performance differences arise. On the other hand, the SOM model does (visually) provide a better understanding of the financial ratings domain.

Conclusion and further research

The SOM display contributes to a better understanding of the underlying domain. The formed clusters are logical groupings of companies, providing a good view on the relationships between the companies in the clusters. Based on these relationships a general assessment of creditworthiness can be given, using specific domain knowledge. This quantitative assessment generally coincides with the quantitative and qualitative assessment as expressed in the S&P ratings of the evaluated companies.

For a more specific view on the creditworthiness of a company we can use the SOM as a classification model. The found clustering is then of less importance; we use the map as a form of non-linear regression. This regression model does not perfectly classify the companies, and this time the model performances are comparable to other, more linear models. One possible explanation could be that the presumed non-linearity is not present in the financial ratios, but is present in the company-specific qualitative information that has not been included in the scope of this research. New research on the credit rating domain could strive to include this kind of information.

Other and future uses of SOM in the financial domain

A reliable signal about the creditworthiness of loans becomes more and more important, due to the rapidly changing market for commercial and government credit. Banks are driven towards desintermediation[6], since the returns from such credits have become marginal and the regulatory environment is getting stricter.

The traditional role of banks in these activities is being gradually taken over by commercial paper markets. Furthermore, in a period of growing economies and lower government debt, public appeal to capital markets is diminishing and alternative investment opportunities may be found in commercial paper. Institutional investors will be the main investors in such markets, but with respect to risk perception there is a big difference between high rated government paper and commercial paper. The asymmetry of information between companies and investors will be much higher.

Traditionally, commercial banks acted as efficient monitors since they had insight in private information that could not directly be transmitted to the public market, for instance competition-sensitive strategic information. In addition, banks had so-called efficiencies of scale to collect and interpret information. Public markets need just as much a reliable signal about what is actually going on. New technology may partly take over such tasks from the credit departments of banks. The effort of advanced data mining techniques in this respect, either by credit rating agencies or directly by institutional investors, will undoubtedly become more and more valuable.

The Self-Organizing Map technique can potentially be used for these kind of analyses, as we have shown in this chapter. The real problem however remains gathering the right data for the case at hand. The sought-after variables are often scarcely available, and the variables that are available often need extensive cleaning and preprocessing before an attempt at extracting knowledge can be made.

This problem is not likely to disappear either in the near or the distant future. Investment opportunities arise, when one has an information advantage on the competition, possibly in the form of a distinctive credit rating classification model. If all the data is freely available to anyone in a good form, it cannot be used anymore to find a distinctive model. As everyone is using the freely available data, the models springing from the data provide the same knowledge to anyone using these models. One should thus incorporate other, more restricted data to improve the model, thereby keeping an edge on the competition.

6 Removing intermediaries from the process.

References

- Bishop, C.M. (1995). Neural Networks for Pattern Recognition. Oxford University Press, New York
- Deboeck, G. (1998). Visual Explorations in Finance with Self-Organizing Maps. Springer Verlag, London
- Eudaptics. (1999). Viscovery SOMine 3.0 User's Manual. http://www.eudaptics.com
- Fayyad, U.M., G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy. (1996). Advances in Knowledge Discovery and Datamining. American Association for Artificial Intelligence Press, Menlo Park, California
- Tan, R.P.G.H. (2000). Credit Rating Prediction Using Self-Organizing Maps. Masters Thesis. Erasmus University, Rotterdam. http://www.eur.nl/few/people/jvandenberg/masters.htm
- Standard & Poor's. (2000). Corporate Ratings Criteria. http://www.standardandpoors.com

High speed quality inspection of potatoes

Jacco C. Noordam[7]

Introduction

Grading and sorting potatoes ensures that the products meet defined grade and quality requirements for sellers and provides expected quality for buyers. Usually, quality sorting is performed by trained human inspectors who assess the potato by inspecting the potato for a particular quality attribute. However, there are some disadvantages in applying human inspectors, such as inconsistency, extensive time to inspect huge volumes and expensive labour costs. Computer vision may improve inspection results and take over the visually intensive inspection work from the human inspector. Computer vision based inspection machines are equipped with cameras which capture an image of the potato. The potato image is digitized by a frame grabber and processed by a computer. The computer segments the potato image into several homogeneous regions (classes) and the number of pixels in each class are counted. Based on the number of pixels within a class and a predefined threshold value, the computer accepts or rejects the potato.

Various studies related to machine vision inspection of potatoes have been reported in literature. Few automated inspection stations for grading of potatoes with machine vision on size and shape are commercially available. However, these systems cannot fulfill the potato industry requirements for high

7 J.C. Noordam, MSc, J.C.Noordam@ato.wag-ur.nl, ATO, Department of Production & Control Systems, Wageningen, The Netherlands

throughput and accuracy. Furthermore, none of these systems is capable of inspecting for size, shape, and multiple external defects.
The objective for this work was to develop a computer vision system to inspect and grade potatoes based on multiple external defects, size and shape. The HIgh-speed QUality Inspection of Potatoes (HIQUIP) system incorporates conveyor belts to transport the potatoes to and from the vision unit. It is assumed dust and dirt are removed before inspection by washing. The system must have a high accuracy and robustness and achieves a minimum capacity of 12 tons per hour.

Overview of the HIQUIP system

The complete potato inspection system consists of a conveyor unit, a vision unit and a rejection unit, all placed in a single line (Figure 1).

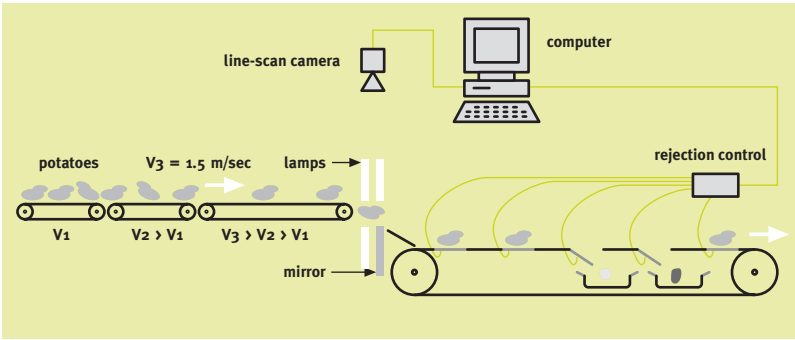


Figure 1
Overview of the HIQUIP system.

Two conveyor belts of the vision unit, placed one after another, transport the potato under the camera for inspection. A digital 3-CCD color line-scan camera scans the narrow gap between the conveyors to achieve in-flight inspection of the potato. To obtain a 360 degree view of the potato, mirrors are placed in the small gap (4 cm) between the conveyors. The lack of product holders and the use of mirrors guarantee a full view of the potato. Three parallel conveyor units are inspected by one camera (left image in Figure 2).

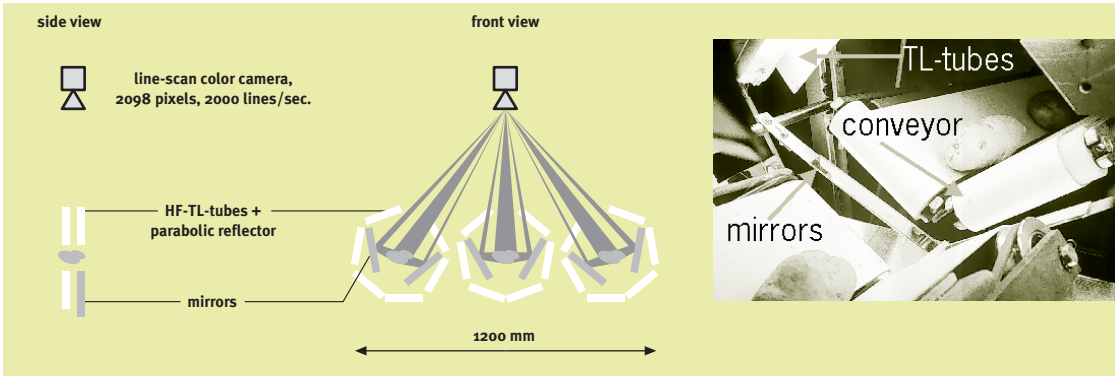


Figure 2
Overview of the camera and lighting set-up (left). Two conveyors with mirrors placed in the gap (right).

Figure 2 (right) shows the two conveyors with the mirrors placed in the gap as the potatoes fly from the upper conveyor (back) to the lower conveyor (front). The left image of Figure 3 shows a top view as the potato passes the gap between the conveyors. The surplus value of the mirrors is immediately shown in the right mirror image, as the crack in the bottom of the potato is still visible in the mirror image. The right image of figure 3 shows the images produced by the camera after image capturing.

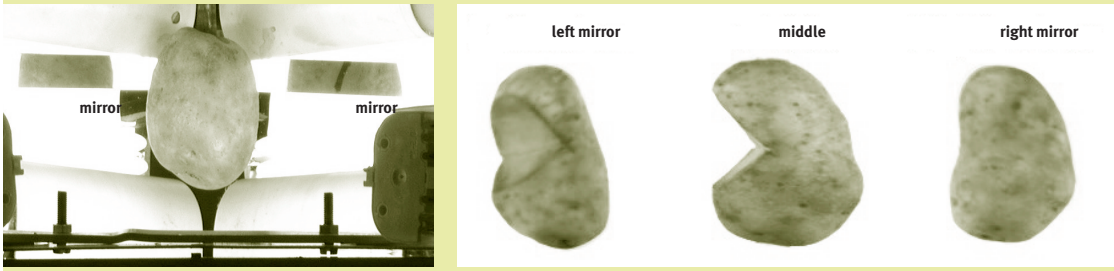


Figure 3
Camera view of the potato as it passes the gap between the conveyors (left). The produced images are sent to the DSP's for processing (right).

To achieve the required capacity of 12 tons per hour each single conveyor unit must process about 10 potatoes per sec at a belt speed of 1,5 m per sec for average sized table potatoes. To detect objects with a size of 1 mm^2, a resolution of 2 pixels per mm is required. To obtain this resolution, the camera must grab at 2,000 lines per sec. To achieve a similar resolution in the direction perpendicular to the movement direction, a resolution of 2,098 pixels is sufficient to cover the width of 1,1 meter for the three conveyor units. A grab frequency of 2,000 lines per sec requires powerful lighting equipment. Therefore, folded small-sized high-frequency TL tubes with parabolic reflectors are used to illuminate the potatoes. Four TL tubes illuminate the bottom part of the potato and six tubes illuminate the upper part of the potato. The illumination amplitude of each individual TL tube is controllable to obtain a homogenous illumination (left image of Figure 2).
The camera grabs continuously and the software detects when a potato passes the gap between the conveyors. Therefore, the camera requires no additional starting signal, when a potato approaches the imaging area.
After inspection, the potato is transported to the rejection unit. The rejection unit consists of individually controlled product holders. Each product holder is kept upwards by electro magnets. Once a potato arrives at the correct rejection lane, the magnets are released and the potato drops at the correct rejection station.

Dedicated hardware is needed for the image processing and classification tasks. A Spectrum Signal PCI-card with 11 SHARC's (Analog Devices ADSP-

21060) Digital Signal Processors (DSP) is responsible for the image acquisition and classification tasks. One DSP communicates with the Host PC and transports the measurement results to the screen for visualization. The other DSP's perform the image acquisition, the color correction and segmentation, image compression and the operations for color and shape classification. An MS-Windows based graphical user interface (GUI) contains various parameter settings and visualization tools for the operator to:

- select different color and shape models for different potato cultivars;
- adjust the margins for the product grading classes;
- learn the HIQUIP system new color defects or new cultivars;
- log or view the history and segmented images of earlier classified potatoes.

Characterization of potato defects

Factors such as size, shape, greening, colored spots, cracks, scab, etc. characterize the final grade of a potato. The reference product expert grades potatoes into four different categories, dependent on the presence of a defect and the area of the defect (Table 1).

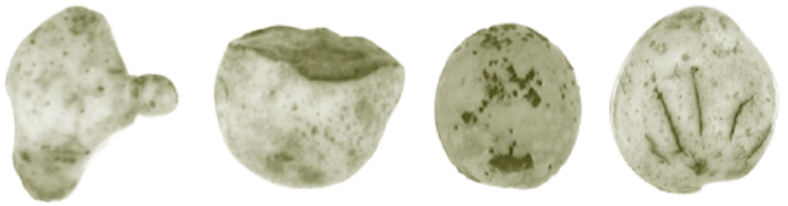
kind of defect	quality classes			
	good potato	**minor defect**	**medium defect**	**major defect**
outward roughness	scale 1	scale 1,5	scale 2,5	scale 3,5
(scab, skin spot, black scurf)	(1-2 spots, ‹ 3,125%)	(2-4 spots, ‹ 6,25%)	(5-10 spots, ‹12,5%)	(20-40 spots, ‹25%)
tuber greening	0-1%	0-2%	2-5%	›5%
pressure spots	not present	0-2%	2-5%	›5%
damaged potatoes	‹ 5 mm²	0-3% (5-10 mm²)	3-8% (10-15 mm²)	›8% (15-20 mm²)
tuber cracks	small cracks	› 0,25cm deep, › 1,5 length tuber not allowed		
misshapen potatoes	not misshapen	misshapen		

Table 2
Grading classes in percentages of total potato area.

Similar diseases on potatoes of different cultivars (scab, skin spot and black scurf) may have a different color due to the underlying skin color of the potato. This requires a different setting for each potato cultivar. Besides the difference in skin color for different cultivars, differences in skin structure and shape are also important features.

For the characterization of defects and diseases, product experts examined five potato cultivars. From each cultivar an image collection of all possible defects was created. This reference database contains more than 1,100 images of five cultivars with the most occurring defects and diseases. The database is used for the development and testing of the algorithms. Four defects on two different potato cultivars are shown in Figure 4.

Figure 4
Four different defects: misshapen, damaged, Rhizoctonia and cracks.



Algorithm design

Color segmentation

The majority of external defects and diseases can be identified purely by color, which makes the classification of pixels into homogeneous regions an important part of the algorithm. Six different color classes are identified: background, potato skin, greening, Rhizoctonia, silver scab, outward roughness. All classes are divided into a number of subclasses, e.g. different background colors belong to the background class and the class good skin exists of dark skin and light skin (Table 2).

Table 2
Classes and subclasses for the image segmentation.

main class	subclass
background	white background, dark background, mirror edge
potato skin	skin light, skin dark
greening	green light, green dark
Rhizotonia	light gray, dark gray, black
silver scab	silver scab
outward roughness	brown light, brown dark

Due to the difference in skin color it is not sufficient to use a single model for different potato cultivars. For each potato cultivar a new color model is created. We use MLF-NN[1] and LDA[2], which are supervised classification routines, so a training set is required to determine the color model. A separate test set is used to test the performance of the classifier.

Discrimination between similar colored objects

There are a number of defects and diseases, which have similar colors. Defects such as cracks and Rhizoctonia both have a black color (Figure 4). To enable discrimination, additional shape features are used. Since cracks and growth cracks appear as more or less elongated, eccentricity is used to identify cracks and growth cracks. A circular object gives an eccentricity of 1; a line shaped object has a higher value. Eccentricity is based on central moments of an object. If the

1 Multilayer Feed Forward Neural Networks. See Section 6.2.8 and the tutorial on NN on the CD-rom.

2 Linear Discriminant Analysis, See Section 6.2.3.

area of an object is above the area-threshold and the eccentricity of the object is above the eccentricity-threshold, the object is classified as a crack. For the detection of growth cracks and normal cracks different values for area and eccentricity threshold are applied. An example is given in Figure 5, where the Bildstar potato contains silver crab, a few small Rhizoctonia spots and a crack. In the segmented image, multiple dark colored objects have been found. Only one of the objects is recognized as a crack by the crack detection procedure, indicated by a black surrounding rectangle.

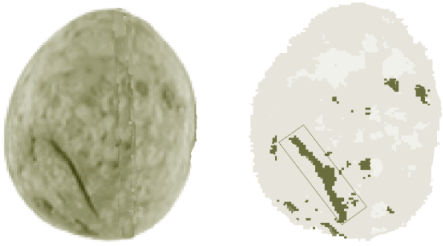


Figure 5
Bildstar image with crack, segmented Bildstar image with detected crack.

Shape classification

For the detection of misshapen potatoes Fourier Descriptors (FD's) are used. They can be made invariant to translation, rotation and scale, which is important in on-line sorting. In the HIQUIP system, FD's and LDA are used to discriminate between good and misshapen potatoes. From each segmented potato image, the boundary is extracted. The one-dimensional boundary is normalized to 256 equidistant points. For each boundary point, the distance to the centroid is calculated. This boundary signature is translated to the Fourier domain and the resulting FD's are the input variables for the LDA. The first 10 harmonics are adequate for representing the shape information of a potato. In the shape classification experiments, different numbers of FD's were used to evaluate the influence of the number of FD's.
A single shape model is not sufficient to segment all potato cultivars into the classes good and misshapen. Good shaped potatoes may vary from round, oval, to extreme oval. Therefore, different shape models are created for different potato cultivars.

Experimental results

Results of LDA and MLF-NN color classification

For five different potato cultivars, a training and test set was created. From the complete set of labeled pixels for each class, 500 pixels were randomly selected

to create the training and test sets. For the MLF-NN, the input RGB values were auto-scaled.
For all cultivars, about 95% of the pixels were correctly classified by the MLF-NN, against 93% for the LDA. Dark colored defects and diseases partly overlap in RGB space and are therefore hard to discriminate by LDA. For this reason the results of MLF-NN are slightly better. However the LDA with a Mahalanobis distance classifier [3] is implemented as segmentation technique in the HIQUIP system as LDA requires no parameter adjustment.

Results of Rhizoctonia recognition by color

In an experiment, the performance of the HIQUIP system is tested for the defect Rhizoctonia. Potatoes of the cultivar Sante were used. Two quality classes are considered; good potatoes and minor defects (Table 2). A product expert inspected and classified each potato beforehand. The good potato class consists of 32 potatoes; the minor defect class consists of 27 potatoes. All potatoes were classified correctly.

Results of shape classification

A set of 40 misshapen potato images from the image database is split in half to create a training set and test set for the misshapen potato class. A set of 130 potato images is split in half to create a training set and test set for the good potato class. LDA with a Mahalanobis distance classifier puts the potatoes into the classes good and misshapen potatoes. The experiment is carried out for 10, 20 and 30 FD's to evaluate the influence of the number of FD's. Using 30 FD's for the boundary description of the potatoes gives the best results, all potatoes were classified correctly. This is to be expected, since extra boundary information is added by using more FD's.

Results of the crack and growth crack detection experiment

To verify the results of the crack detection procedure, 20 potato images with cracks and 75 potato images without cracks are selected from the image database. The 75 potato images without cracks are covered with similar colored defects of the classes Rhizoctonia and outward roughness. After the color segmentation, eccentricity and area are calculated for all objects of the classes Rhizoctonia and outward roughness. All cracks were recognized correctly.
To verify the results of the growth crack procedure, 64 potato images of the Bintje cultivar are selected from the potato image database, from which 9 potato images were labeled as growth crack. The remaining 55 potato images are covered with the similar colored defect common scab. After the color segmentation, eccentricity and area are calculated for all objects of the class common scab. The HIQUIP system recognized 97% of the potatoes with growth cracks.

3 A classifier used to express the degree of similarity of objects, with some advantages compared to Euclidian distance.

Conclusion

A high-speed color vision system for the inspection and grading of potatoes has been developed. The HIQUIP system grades potatoes on shape, size, cracks, growth cracks, and color defects such as greening, common scab, silver scab and Rhizoctonia. A 3-CCD color line-scan camera inspects three parallel conveyor lanes, moving at 1,5 m per sec, to achieve the required capacity of 12 tons per hour. Mirrors placed in the narrow gap between the two conveyors guarantee a full 360-degree view of the potato. Multiple folded small-sized high frequency TL tubes with parabolic reflectors are sufficient to get powerful lighting. The PC-based PCI board with 11 Digital Signal Processors performs the image processing and classification tasks with a speed of 50 potatoes per sec.
LDA and a Mahalanobis distance classifier classify RGB pixels in six different color classes. Pixel classification experiments show classification rates above 90% for five potato cultivars. Color defects like greening and scurf can also be recognized with high accuracy. With area and eccentricity as additional features cracks and growth cracks can be identified from other defects with similar color. The Fourier-based shape classification procedure is powerful in detecting misshapen potatoes.
More tests are necessary to evaluate the performance for other defects and different cultivars. The capacity of 12 tons per hour and the reported classification results indicate that the HIQUIP system can fulfill the demands of the potato packing industry.

Future research

An important future research item is the improvement of the self-learning potential of the color classification in the HIQUIP system. As the sorting of new potato cultivars requires a new training set, a more or less automatic adaptation of the system to the new colors of the potato cultivar is highly desirable. Current training set selection is based on mouse selection in an image that contains a representative potato with a new defect. This technique has a few drawbacks:

- It is very time-consuming to select a learning set with enough pixels, especially when a lot of different classes are involved.
- There is a reasonable chance that a pixel or a group of pixels will be selected which does not belong to the same class as the already selected pixels. In that case, the poorly selected pixels will contaminate the class and will undoubtedly influence the segmentation results.
- Mostly, a single image is used for the creation of a training set. A non-representative image leads to a bad training set and most likely, this results in a bad segmentation. It is preferable to use a representative number of images that form a good reflection of the real world problem.

A solution to the problems above is an algorithm which automatically selects (unsupervised) a training set from multiple images. The automatically created training set can be used to train the segmentation routine (LDA in the HIQUIP system) and segment the images in representative regions. Such a training set selection technique can also be used in inspection systems for the quality inspection of pot plants, French fries and meat.
An important problem to be solved in automatic training set selection algorithms is the required high level of accuracy and robustness. The performance of the HIQUIP systems strongly depends on the performance of the color segmentation and thus on the quality of the training set.
Another future research item will be the utilization of multi-spectral cameras to obtain better discrimination between similar colored defects. These cameras are able to capture images in the UV or Near InfraRed part of the electromagnetic spectrum. Defects and diseases which are hard to discriminate or even invisible in the visible spectrum might become visible in other areas of the spectrum. As these cameras produce a huge amount of data, the use on a priori product information and the use of intelligent data reduction and extraction techniques will be an important issue.

References

- Noordam, J.C., G. W. Otten, A.J.M. Timmermans, B. van Zwol. (2000). High-Speed Potato Grading and Quality Inspection Based on a Color Vision System. Presented at SPIE, Machine Vision and its Applications. San Jose, California, USA
- Noordam, J.C., A.J.M. Timmermans, G. W. Otten, B. van Zwol. (2000). A Color Vision System for High-Speed Sorting of Potatoes. Presented at Agricultural Engineering 2000. University of Warwick, Warwick, England
- Tan, R.P.G.H. Credit Rating Prediction Using Self-Organizing Maps

3.2.3 Detecting

Introduction

Defined earlier in this part as 'being able to detect a deviating state that is considered relevant', detection is mainly a warning signal that triggers action. Techniques used vary from standard statistical techniques to classification and advanced image and pattern discovery and recognition.

Examples of applications include intrusion detection in a network, detection of fraudulent phone use or suspicious money-transfer patterns. By video surveillance it has become possible to detect suspicious behavior by car thieves or burglars in car parks. They tend to exhibit certain patterns that deviate from regular users of a parking. See also Chapter 5.5, Multimedia mining.

A general approach for detecting deviations is given in [Arning, 1998]. Examples in fraud detection are given in the next article. Pro's and cons of data surveillance are discussed in [Clarke, 1988]. The second case in this section, Detecting irregularities in waste transport, is an example that illustrates a wide variety in data sources.

Detecting suspicious behavior

Christopher Westphal[1]

Introduction

Catching a thief, spy, money launderer, insider trader, tax evader, narcotics trafficker or anyone exhibiting patterns of 'non-compliant' behavior can be effectively performed using a combination of strategic (pro-active) and tactical (reactive) analytical techniques. The exact definition of a non-compliant pattern will vary according to the activities occurring within the domain for which it is being applied. In many instances, the methods used to identify a pattern will change depending on the situation — the challenge to the investigator is to discover what associations in the data constitute a reliable and valid pattern. This paper overviews several methodologies used to identify patterns of non-compliant behavior across multiple and disparate data sets (for instance financial crimes). Real world examples utilizing an operational visual data mining application targeted at money laundering operations are presented to show exactly what conditions were used to identify different types of behaviors based on a set of transactional events (e.g. the activities).

1 C.R. Westphal, westphal@visualanalytics.com, Visual Analytics Incorporated, Poolesville, MD 20837, USA, http://www.visualanalytics.com

Data

In relational data management systems, or any electronic media for that matter, data is usually taken at a face value, that is, without trying to understand much

about how and what is being represented. We typically gather information and store it into buckets (database tables and fields), so we can recall it at some future point, hoping that it may be useful or provide value to an ongoing case or investigation.

In the law enforcement community, enormous volumes of information are being collected and stored in this fashion without regard to how it will ever be used. Thus, huge repositories of information exist which most likely may never be effectively utilized. This is due in part, because there are no established means (technologies or methodologies) by which to easily understand what is contained in them. Thankfully, initiatives supported through government sponsorships and commercial initiatives have provided new approaches to help law enforcement organizations better understand, manage, and present their data. Furthermore, these advances are applicable across a wide range of functional areas.

Understanding data

The goal of many law enforcement investigations is to find associations or establish relationships among the targets or suspect entities. Associations can be made not only based on the name of a person or organization, but also through addresses, identification numbers (including licenses, passport, or social security numbers), telephones, vehicles, accounts, and most importantly activities. Activities represent the phone calls, border crossings, cash deposits, and meeting contacts associated with the illicit conduct being analyzed. A single occurrence of any activity, in and of itself, is not ordinarily considered of interest. However, when taken as a collective whole, a much different set of patterns emerges.

The behavior associated with activities (also called events) can be exposed using several different approaches. The goal is to take advantage of the similarities to understand the underlying behaviors associated with the activities being reviewed. Depending on the nature of the application, the behaviors will indicate different types of patterns that, once known, can then be identified and exploited according to the specific needs of the investigative agency. An analysis of activities can expose general trends or very specific patterns. Whether you are looking for temporal sequences or spatial dependencies, all information is contained within the activities themselves.

Descriptive data models

Virtually all data modeled in an analysis may be characterized as either descriptive or transactional. Descriptive data describe objects such as people, places, or things via attributable values (such as the location of a burglary, the subscriber of a phone, the color of a vehicle, etc.). The attributes of any descriptive

object are unique to the object and are also considered a one-to-one mapping — an attribute should only support a single value.

Descriptive data typically represent 'state-based' knowledge that is considered true or believed until replaced by a different value. Once replaced, the old value is usually not maintained nor is it used within any subsequent analyses. Addresses, persons, accounts, vehicles and amounts of money are examples of descriptive data elements.

Descriptive data tend to represent information such as organizational structures, credit report headers, driver's licenses, last known addresses, parole terms, account ownership, and so on. When various classes of declarative information are used in a model, one can show that relationships exist among them. The models are the explicit translations or mappings of the raw data into a primary object representation (what are objects, what are attributes, and what are defined to be the relationships).

For example, you might have a model in which there is a class of objects corresponding to people and another class corresponding to vehicles with relationships between them indicating who owns which vehicles. Descriptive models tend to outline the overall structure of the relationships between the different objects contained within a data set. Descriptive data are useful for looking at networks and frequencies of connections, but are very difficult to use if the goal is to describe behaviors or events.

The networks that form from the definition of descriptive data can be used to expose a wide range of important data patterns. Since the goal of most law enforcement investigations is to establish relationships, descriptive data is a perfect representation format for supporting these tasks. The following depicts but a few of the structures that can be derived from the use of declarative data sources. Of course their exact meanings and usage will be dependent on the context in which they are used.

- *Articulation points* — look for bottlenecks where one particular entity connects two or more subnetworks. These entities represent potential vulnerabilities and can be exploited, targeted, or used to the benefit(s) of the investigating organization.
- *Missing links* — expose entities that are detached or unconnected from the main network structure. The investigator needs to determine why the entities are isolated and if there are missing data that would tie them in with the rest of the data.
- *Discrete networks* — identify all the isolated subnetworks contained in the data. These groupings can be used to focus resources and help expose the extent of an investigation.
- *Strong/weak linkages* — see the strength of relationships within the network. Those entities with multiple (strong) relationships can be subject to transactional (behavioral) analyses.

– *Pathway analysis* — determine if a series of linkages will connect a tuple[2] of entities including the shortest path. Finding indirect relationships among suspects is vital in law enforcement applications.
– *Commonality* — look for entities connected to or that share common elements. These situations help identify additional targets as well as show the overlap of known resources.

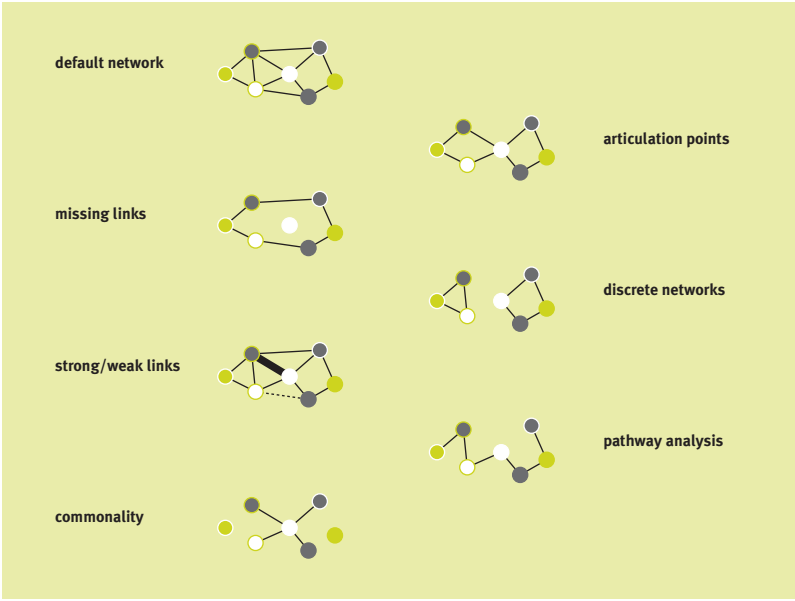


Figure 1
Network patterns.

Figure 1 shows a small network diagram for each one of these situations. Keep in mind that there are many different permutations and variations on the descriptive data pattern theme. The investigative domain largely defines the interpretation. For example, the patterns identified for telephone toll analysis will be different than those identified for narcotics investigations.

Transactional data models

Transactional data, in contrast to descriptive data, contain episodic information about time and place of events. Much of the information within database systems used by law enforcement represent transactions of some sort. The structure of transactions remains fairly static, that is, the content or values stored in transactional structures is what varies among instances. In general, transactional data contain a date/time component that can be used as a primary key to distinguish each discrete transaction (the transaction itself is what is unique).
To get a feel for transactional representations, consider the use of your telephone. No matter how many transactions (phone calls) are recorded for your telephone, the structure of the representation remains consistent. There is always a destination number, time, duration, and a date. What makes each tele-

2 A data object (row) containing two or more components.

phone call unique are the values which are applied to each of these attributes for every transaction. Even when you call the same telephone time and time again, each transaction will be unique just from the fact that it occurs at a different time and date. Many domains are heavily based on the use of transactional data, such as narcotics, electronic messaging, terrorist activities, telecommunication, contacts scheduling and money laundering.
Since every transaction can be distinguished from every other by its assigned values you can start to perform an analysis to look for explicit transactional patterns and related behaviors.
In a transactional model, you can typically use links between object classes to represent traits or conditions of the event contained within the transaction. Thus, all of the conditions associated with the event can be applied to the link, since they were derived as a result of the event. What this means is that any attribute applied to any link generated between any pair of objects derived using a transactional model can justifiably support any of the information used to describe the transaction.

Is it not true that the amount, time, or date of a financial transaction can be used as an attribute to describe the link created between the transactor and the account as well as the link between the address and the business? Also, the existence of many links between two objects indicates that many separate transactions occurred between them and each is represented in the data set.
In a descriptive model, on the other hand, each of the links can have its own unique value used to describe the relationship between object classes. Thus, an individual person in a 'people' object class might be connected to an 'address' in another class, because the information was listed in a credit application. However, that same individual can have a link with a particular automobile in a 'vehicles' object class for a completely different reason, defined by a completely different set of conditions. This flexibility occurs in a descriptive model, because the information that is used to specify the conditions of the relationship is distinct to the objects of interest. In a transactional model, the conditions represented in any relationship are generated from the transaction record and so are more focused on the event itself.

Thus, we could immediately look at the results to determine whether there are any trends of interest with regard to these behavioral questions. You should note that although the transactional model is the most powerful alternative for examining behavior, it comes with a price. Transactional models often involve the proliferation of large numbers of objects and links that must be managed. On the whole, there are no tremendous differences in the qualitative nature of individual transactions. For example, all phone calls share certain traits or attributes in common. What is important in the analytical environment is that

this episodic information can be used to distinguish one transaction from another. That is, you may not be so interested in what happened (since all events are fairly similar to one another) as you are in when, where and how often it happened. In contrast, semantic or descriptive representations may or may not contain the same attributes. Patterns of interest within descriptive knowledge structures often center on similarities and differences in these attributes.

When an investigation is initiated, you will need to decide whether you are dealing with descriptive data, transactional data, or a combination of both. The type of data will determine the analytical models that can be used. If your data are descriptive, your analyses will be confined to the use of descriptive models. Transactional data, on the other hand, may be represented in either descriptive or transactional models. The types of results that you will be able to glean from your analyses depend largely on the kind of model you select.

Show me the money

Financial crimes are a lucrative business for both criminals and the law enforcement organizations that pursue them. There has been a good deal of literature in recent times describing new and innovative ways of detecting money laundering operations.

In the US, regulations imposed on banks and other financial institutions were designed to generate a trail that could be used to track the path of money as it moves through the system. There are several different types of forms that must be filed when a cash transaction of currency for more than $10,000 is conducted. Banking related industries are required to file an IRS Form 4789 - Currency Transaction Report (CTR).

Businesses such as car dealerships, jewelry stores, and any other retailer dealing with large dollar merchandise are required to file IRS Form 8300 - Report of Cash Payments over $10,000 Received in a Trade or Business. Additionally, casinos need to file IRS Form 8362 - Currency Transaction Report by Casinos, and any international travelers file Customs Form 4790 - Report of International Transportation of Currency or Monetary Instruments.

Finally, those people who maintain foreign bank accounts are required to file Treasury Form TD F 90-22.1 - Report of Foreign Bank and Financial Accounts (FBAR). There is even a special form called a Suspicious Activity Report (SAR) TD F 90-22.47 that is submitted by financial institutions for any suspicious financial transactions, including those under the $10,000 trigger.

Well over 100 million stored CTR forms are stored in various database systems throughout U.S. state and federal agencies.

To effectively understand all of the behaviors associated with this type of data, its format must be understood. Naturally, the information contained on a CTR can be viewed using either a descriptive or a transactional data model. Using a descriptive model, there are all sorts of information that can be exposed. The primary data objects used in this model typically represent people, organizations, identification numbers, accounts, banks, and tellers. From here a variety of structural patterns can then be viewed to look for various financial crimes. Depending on the agency, the structures will expose different patterns. Figure 2 shows a set of objects where a central entity has connections to two different entities. Their interpretation is based on the context in which they are being analyzed.

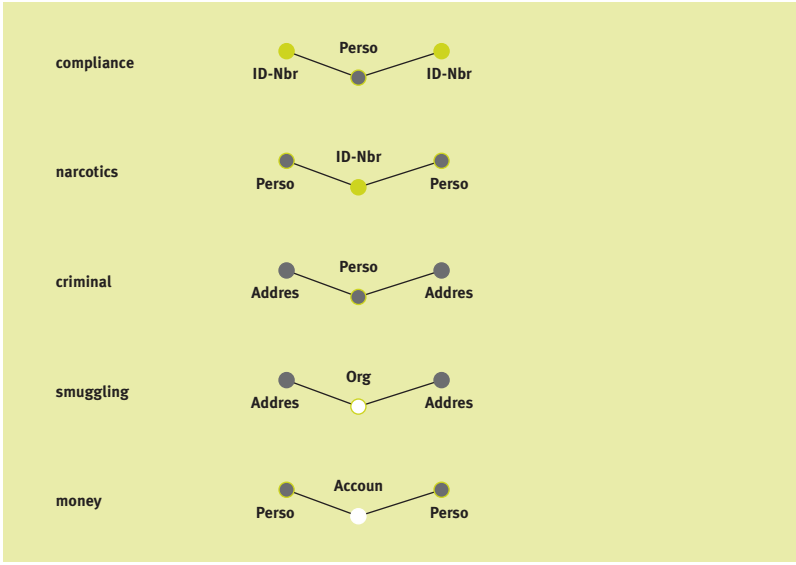


Figure 2
Descriptive CTR models.

For example, when one person is connected to two different social security numbers it may indicate a non-compliant tax situation to the IRS. Alternatively, two people connected to a single id-number may show deceitful behavior where they are trying to avoid detection. Multiple addresses for any one person might lead investigators to suspect illicit activities are ongoing. The combinations are virtually endless.

When looking at the transactional CTR models, for brevity, lets assume that the only information to be modeled represents the subjects conducting the transactions (the transactors) and the CTRs (the transactions) themselves. Remember that the attributable information that is contained on the transaction itself (the CTR) can be used to expose the underlying behavior associated with moving the money. The following represents only some of the details required for filling out a CTR:

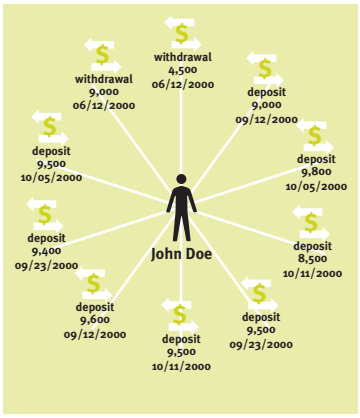
- account type;
- amount cash in/out;
- bank location;
- bank name;
- date of transaction;
- occupation type;
- teller/official name;
- transactor address;
- transactor id;
- transactor name.

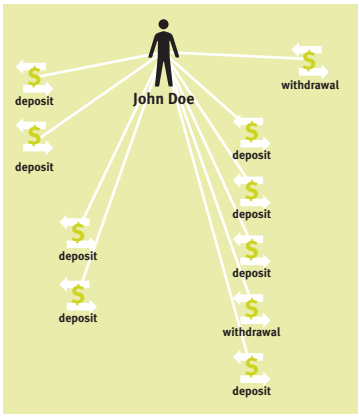
The goal is to provide a well-integrated picture of all of the known cash transactions for a transactor, when money is moved in or out of a financial institution. Once a pattern has been identified, law enforcement can take appropriate actions to prevent, circumvent, or interdict the targets based on their known behavior.

Transactional examples

To make better sense out of what the transactional patterns would look like in a real world financial crime application, several hypothetical examples have been created to demonstrate the concepts. Typically no single display indicates there is something suspicious going on, the investigators must look at all the dimensions and come up with an overall evaluation of the situation.
The diagrams generated for these examples were created using the VisuaLinks™ data visualization software. However, it is the concepts and methods that are important to understand, and not the specific presentation technique.
Figure 3 shows a transactor who has been involved with a number of CTRs. The transactor appears at the center and is linked to each of the transactions that are shown in the display. The links are created to make the relationships explicit even though the value of the transactor is an attributable value in each of the CTR objects. The links make the display look more consistent and are useful, when more than one transactor is being investigated.

For these purposes, a highlight-color for any CTR object can be defined to reflect certain values. In this case it was set to show, if this transactor has ever been previously reviewed by any other law enforcement agency. As it turns out, there are no prior reviews, which potentially means he maintains a well-established pattern (e.g. relationship) with the banks. Of course this also means he may be very good at avoiding these situations or perhaps the bank has already been subject to infiltration or corruption.





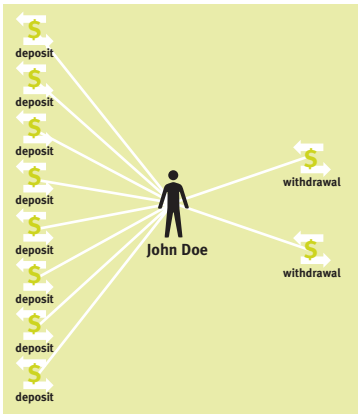


Figure 3 (left)
No prior audits/reviews.

Figure 4 (middle)
Multiple branch activity.

Figure 5 (right)
Bias for a teller.

Figure 4 shows another example of the same transactor where the CTRs have been clustered (grouped) according to the particular branch where the transaction took place. As can be seen, there are now four discrete clusters for the CTRs. The majority of transactions have taken place at one particular branch as is seen by the larger cluster. However, the presence of the other three clusters may indicate there is potentially an ongoing 'structuring' pattern.

It is illegal to 'structure' a series of transactions to avoid the $10,000 threshold. For example, you cannot knowingly manipulate a financial institution into failing to file a CTR report by making three deposits of $9,000 instead of one $27,000 deposit. Using multiple branches to move large cash deposits is not consistent with legitimate behavior.
Figure 5 represents a situation where the CTRs have been clustered based on the teller who has performed the transaction on behalf of the filing institution (e.g. the bank). In this example there is an overwhelming bias for a specific teller.

A possible explanation for this situation might include that the bank has one designated official who authorizes these types of transactions and therefore, by default, his or her name or id-number is listed on all the CTR forms. Much more likely is that there is a collusion with one of the tellers. This is also reflected in the link colors (not shown in the B&W diagrams) where the teller is shown to have operated at more than one branch location, when the transactor has made deposits. Notice that this teller performs all deposits and another teller performs all withdrawals.

This type of diagram can be used across a wide range of the attributable values present on the CTR form. Figure 6 depicts a slightly different display configuration where the CTRs are laid out according to the date when they occurred. This format is considered to be an 'absolute' placement.

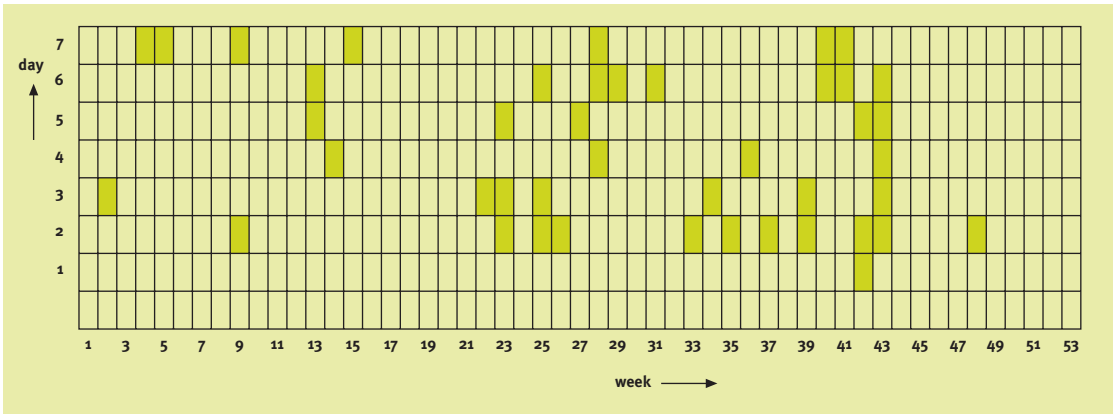


Figure 6
Irregular filing dates.

The placement of CTRs according to their dates can be extremely powerful, especially when dealing with larger quantities of information. Patterns are more readily detected using this type of display, since there are so many dimensions being displayed at one time.

Within this display the CTRs are presented using a 7x52 placement algorithm. Dates associated with a CTR can be used to determine the day-of-week (1-7) as well as the week-of-year (1-52). The resulting matrix definitively presents the temporal filing patterns associated with the transactor. It is very clear due to the time-gaps that the transactor does not follow a regular pattern, which implies there is no steady income, nor is the exhibited pattern consistent for any known occupation codes (a required field on the form).

There are all sorts of variations on this theme, especially when dealing with times and dates. Since transactional patterns are structurally consistent, the same type(s) of patterns can be derived from telephones, border crossings, travel events, email, web site visits (cyber crime), or just about anything else with a temporal value.

Ideally when dealing with financial crimes, the ultimate goal is to seize the assets associated with the money laundering operations. Thus, if a temporal pattern can be identified and confirmed, then the law enforcement agency has a better chance of actually obtaining a 'cash' assets forfeiture, because they can predict when the funds are going to be moved or when the accounts are full.

Conclusions

There are good people – and there are bad people. The bad people cost the good people a significant amount of monetary and resource losses (measured in billions of dollars) through the liabilities incurred from fraud, theft, espionage, embezzlement, public corruption, and proliferation.

In many cases the malpractice and malfeasance succeed, because people do not know how to interpret their data sets or recognize the telltale symptoms.

The majority of wrongdoing is carried out in a large number of relatively small exchanges. A large percentage of crimes such as money laundering are perpetrated through a series of frequent transactions with relatively small amounts of money being processed on any one occasion. This sort of activity is of course subtle and not directly detectable through usual methods of oversight. To catch a thief, or any wrongdoer for that matter, one must lock onto a behavior pattern. Data mining and visualization approaches can be applied to these sorts of problems with great success at relatively low cost.

Detecting irregularities in waste transport

Jochen van der Wal[3]

Situation and problem

Public opinion demands measures against environmental pollution, as a healthy environment is very important for the common welfare of the population. Politics insist that more attention should be paid to maintaining the environmental laws which prevent environmental pollution. Some of these laws could be directly extended to traffic law enforcement. For example, the police should be required to supervise the transportation of toxic waste and other dangerous goods. A lot of these hazardous goods are transported by road.

This case uses the transportation of waste and dangerous goods as an example of the way in which the police can use data mining to achieve more efficient management of traffic law enforcement. A system called 'Well-targeted traffic supervision' is already in use by the Dutch police forces. A special group focuses on targeted traffic supervision to reduce accidents and stimulate traffic throughput. The idea is to target only the likely offenders for further investigation and to leave innocent road users unaffected. Police resources are very limited, so these resources should be used in an intelligent way. Data mining and related techniques can help in managing these resources. This article discusses ways to detect possible suspicious activities in industrial waste transport through a combination of data from several sources. The analysis itself is not described in detail.

Characteristics of the available data

Different data sources can be used. Some of these data sources could be useful for data mining purposes. The useful sources for this context are summarized in Table 1.

3 J. van der Wal, MSc, Dutch National Police Agency, Criminal Investigation Department. Research and Development, Driebergen Rijsenburg, The Netherlands.

data source	information	form	digital interface possible
magnetic road loop (Monica)	footprint, velocity, and movement of the vehicle	digital	yes
video observation system	number plate, type and color of vehicle	image/video	no
recorded law enforcement facts	type, place and time of the offence. The persons and vehicle concerned	digital	yes
cargo and transport information	cargo, destination of vehicle, vehicle information, and vehicle owner information	paper	no
vehicle information	number plate, vehicle type, carrying capacity, owner, insurance information	digital	yes

Table 1
Data sources. The table above does not represent the complete picture of all available data sources, only the sources needed in this case are mentioned.

At the moment not all mentioned sources are available and accessible in digital formats. Naturally, privacy issues become important, when different sources are combined to extract conclusions from the combined data with data mining tools. However, the discussion of these issues is beyond the scope of this article. Privacy issues are discussed extensively in Part 4.

Choice of techniques and motivation

The available footprints from the magnetic road loops are the starting point of the observations. These footprints are already used to classify the passing vehicles, to measure speed and to detect vehicle movements. The possible vehicle classifications are:

- motorcycle;
- cars with and without a trailer;
- truck with and without a trailer.

The footprint contains more information than has already been mentioned. A closer look provides information on the type of vehicle passing the magnetic road loop. Theoretically speaking, the changes in magnetic field induced in the road loop for (for example) a normal VW Golf and a VW Golf GTI differ from each other. When the magnetic fingerprint of a car type is different from other car types, identifying the car type becomes possible. There are limits however: of course it is theoretically impossible to detect a difference between a yellow and a green VW Golf of exactly the same type, in spite of the fact that we are talking about completely different cars with different owners and different number plates. To identify a specific car (and the route the car takes), additional information is needed. Information of vehicle movements from the traversed magnetic loop system could enhance the information in such a way that it becomes possible to detect a specific car. Information from the video observation system can be used to check the identity of the vehicle on a periodical basis. To do so, automated detection and identification of the vehicle is necessary. The informa-

tion from the video observation system should be matched with the information from the magnetic loop system to obtain greater certainty about the identity of the vehicle.
The result of all this analysis is the description of the route of the vehicle. The actual, driven route obtained can be instantly matched with the expected route, defined in the cargo and transport information. An automatic system can detect a difference between the actual route and the expected route in real time. The police can use this detected deviation and react on it (bear in mind that we are talking of transport of hazardous chemicals and waste). Standard statistical methods are used for this detection task. Furthermore, the detection of the actual weather circumstances with the knowledge of announced transport of dangerous goods can result in an offence, for example driving dangerous goods in fog. It is of importance also to detect transports of dangerous goods through tunnels which are restricted for dangerous goods transports.

Implementation issues

The extraction of a vehicle classification from the footprint information of the magnetic road loops is already in use. Refinement of this information into extraction of the vehicle type is still imaginable, but identifying specific vehicles from nothing but the footprint is theoretically impossible. To identify a specific vehicle, other information such as the movement of the vehicle over several magnetic road loops and identification from the video observation system is necessary. For this purpose the magnetic road loop system (Monica) and the automated video observation system have to be linked together. Furthermore, the number plate (and possibly the color and shape) of the vehicle have to be recognized from the image material of the video observation system.

This raises issues with a more political nature. All the information should be available, real-time and in a well-defined digital format. All the different parties have to cooperate in information exchange. Furthermore, the privacy aspects of combining different data sources need a thorough review.

Future possibilities

In the near future it should be possible to detect any deviation of regular road behavior. Especially for transportation of waste and dangerous goods like gas or toxic substances, prevention of any dangerous situation is vital. Police actions could help to stimulate the drivers to behave according to guidelines. For maximum efficiency, these actions should only take place, when the driver makes a serious mistake: no police action is needed when road behavior is considered regular. When waste is moved from one place to another, the police need to know if the transport is legal or not. Only irregularities in transportation of waste are interesting for the law enforcement task. Examples of interesting

situations are when a driver dumps toxic waste or picks up illegal waste to dump it together with legal waste.

Recently, a discussion about electronic vehicle identification has started in various countries. This system uses an electronic 'number plate' to identify a vehicle. An electronic number plate has an advantage over a normal metal plate, since the normal plate is not always visible. Dirt, weather circumstances and obstructions hinder visual inspection of the plate. Identification of the electronic number plate is possible through fog and small obstructions, increasing the chance of identifying a vehicle.

The introduction of a large number of roadside systems to pick-up the signal from the electronic number plate would create the infrastructure to trace vehicles (Electronic Vehicle Identification, EVI). This would be an easier and more secure way to trace transportations of waste and toxic/dangerous goods, for example. The communication between the electronic number plate and the roadside system could even be used to transport a variety of information available in the vehicle. Summarizing, the total information that could be available in the near future is listed in Figure 1. Various levels of privacy could be used, from high (regular passenger cars) to low (dangerous goods and waste). Obviously, with this higher level of information, detecting irregularities would become easier.

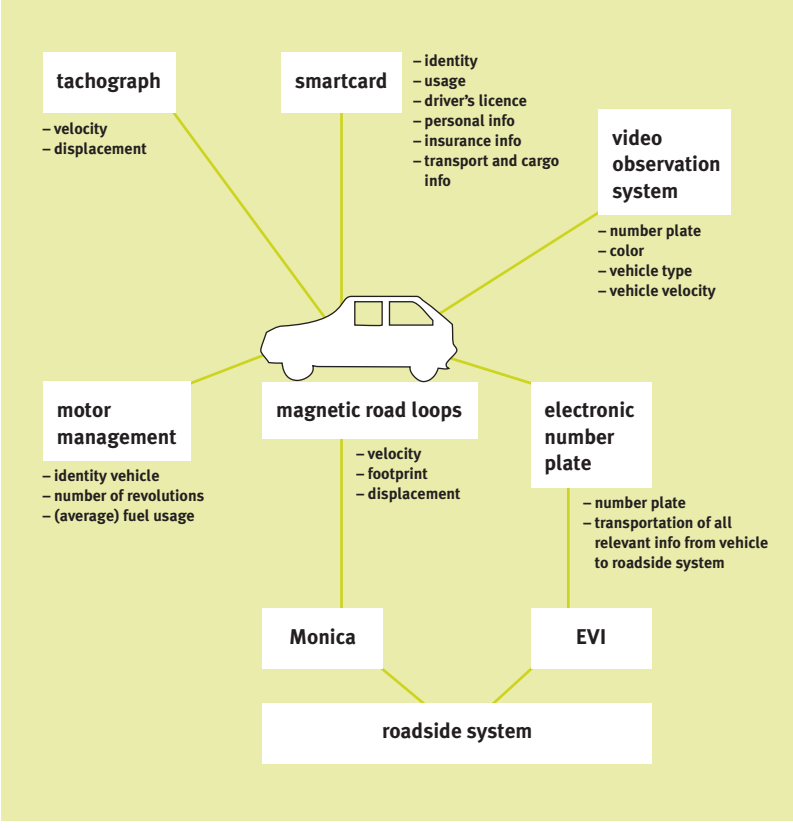


Figure 1
Overview of the different information sources available in a vehicle.

References

- Arning, A., R. Agrawal, P. Raghavan. (1998). A Linear Method for Deviation Detection in Large Databases. IBM German Software Development Laboratory, Boeblingen, Germany, IBM Almaden Research Center, San Jose, California, USA
- Clarke, R.A. (1988). Information Technology and Dataveillance. Communications of the ACM May **37** (5)
- Westphal, C., R. Arndt, M. Kling. (1999). To Catch A Thief...Visual Analytics Incorporated. Bethesda, USA

3.2.4 Modeling

In Chapter 3.1 we described modeling as 'generating an abstract description of (a part of) reality'. Although the model is often a starting point for further analysis or even prediction of values, in this paragraph we will focus on the cases where we are interested primarily in the model itself, as a way to understand a process or working principle. With this understanding we can develop better controls, regulating mechanisms, products or procedures.

Data mining in rehabilitation and ergonomics

Chris T.M. Baten[1]

Introduction

A central issue in physical rehabilitation is the assessment of the human motor function. More and more, recent technological developments facilitate the assessment of the subjects motor function in more natural environments outside the laboratory.

Miniaturization and technological evolution of inertial movement sensors and surface electromyography sensors enable accurate monitoring over longer periods of time of posture, movements, muscle activation patterns, muscle condition and even joint moments and forces [Baten, 1996] (Figure 1).

When combined with global subject behavior and activity information this generates a wealth of information for evaluating recovery and other rehabilitation processes.

Very similar technologies are developed and are already partly deployed in ergonomics to assess physical load exposure in order to predict and evaluate work related health risks [Baten, 2000; Baten, 1995]. This is done both with healthy subjects in a preventive fashion and with subjects who have already developed complaints. Here the fields of ergonomics and rehabilitation blend into each other.

In both fields data mining will play an important role transforming the tremendous amount of data, typically generated in ambulatory assessment, into knowledge and models of the influences on the human motor function. This will lead to sensible clinical and ergonomic diagnoses, predictions and decisions.

Ambulatory assessment of work related back pain risk

As a typical example of the state of the art and future challenges for data mining in the fields of rehabilitation and ergonomics a method for ambulatory assessment of physical load exposure is presented. This method was developed over the last 6 years under the acronym of 'AMBER' and is nearing the phase of wide

1 C.T.M. Baten, MSc, Chris.T.M.Baten@RRD.nl, Roessingh Research and Development, Enschede, The Netherlands

spread practical application enabling large scale data mining into predictors of work related complaints [Baten, 2000].

The method estimates physical back load exposure in terms of net moments and forces in the lower spine from posture, movement and muscle activation data assessed with miniature inertial sensors and surface EMG [2] sensors. Contributions of the trunk and head masses to the physical load exposure are assessed from absolute 3D trunk orientation, trunk angular velocity, angular acceleration and linear acceleration, all assessed through one to three inertial movement sensor modules applying more or less standard mechanics [Baten, a, Baten, b].

The contribution of the arm masses and manual material handling are estimated from trunk muscle activation patterns in conjunction with trunk kinematics. For this the relationship between EMG plus the kinematical input data and the net moment contribution output data needs to be established (or calibrated) anew in each individual subject assessment. To enable practical application a fast self learning method was developed applying artificial neural network technology [Baten, 1995; Baten, 2001], see Section 6.2.8, Neural networks. Additional synchronously recorded video footage and observation data complete the data set in order to facilitate practical interpretation.

Summarizing the presented method for ambulatory assessment generates a large set of types of data, very different in nature, from which intelligent and extreme efficient information compression is required in order to develop a risk profile for work related back complaints.

In Figure 1 3D trunk posture and movement data are estimated combining data of 3 piezo-resistive linear accelerometers and 3 solid state gyroscopic angular rate sensors, mounted with all the relevant electronics in modules sizing 3 by 3 by 5 cm.

Figure 1
Miniaturization and technological evolution of inertial movement sensors and surface electromyography sensors enable accurate monitoring over longer periods of time of posture, movements, muscle activation patterns, muscle condition and even joint moments and forces.

2 Electromyogram, measuring electrical activity in muscles.

Muscle activity is recorded through surface EMG techniques and global posture, movement and activities are assessed through wireless time-synchronized video and multi-moment observational methods.
Control over the recording session, data storage and wireless synchronization with a portable DV camcorder and observation hard- and software are all performed by a portable data acquisition system (right). The subject can move freely and without interference from the recording equipment and can be monitored during natural activities like, in pig-tending, trying to catch a piglet hiding behind his mother (top left) or, as a train arranger, changing tracks (bottom left).

Characterization of available data

A typical session generates the following set of data. First there are 3D posture and movement data. For each sample instant (100-1,00 samples per second) these include matrices for 3D absolute orientation, 3 element vectors for 3D angular rate, angular acceleration and linear acceleration. These data can be present for the trunk of a single body segment or for each vertebra separately. Then there are scalar signals for the amplitude envelope for all the involved muscles. There are 4 to 10 muscles involved.
The 3D net lower spinal moment and force are both 3-element vector signals. These are at least available for one intervertebral body, but when the trunk is modeled in separate segments also for 2 or 3 other intervertebral bodies. In addition, net moments are always separately available for the trunk, for the arms plus manual material handling. Furthermore, they can be split up into components related to different generating forces e.g. the component generated only by gravity versus the component generated by accelerating the upper body by stretching the legs.
For global reference of the activity of the subject wireless synchronized DV-video data are available, typically at a frame rate lower than that of the physiological data (25 to 50 fps). These will be stored in a MPEG format and analyzed.
For reference of task details and task handling of the subject behavioral observation data are available. These take the form of a series of channels containing event timing data or task element state data. Events have zero duration and task elements states have nominal data values during a certain time. Figure 2 gives an example of a typical data set.
Methods to improve the visual grasp of the movement and posture with a photo-realistic virtual spine are under study.

State of the art of data mining

Very little is yet known about the quantitative relationships between the typically biomechanical data collected with ambulatory methods and the actual risks to work related complaints. Lots of indications are reported in many studies

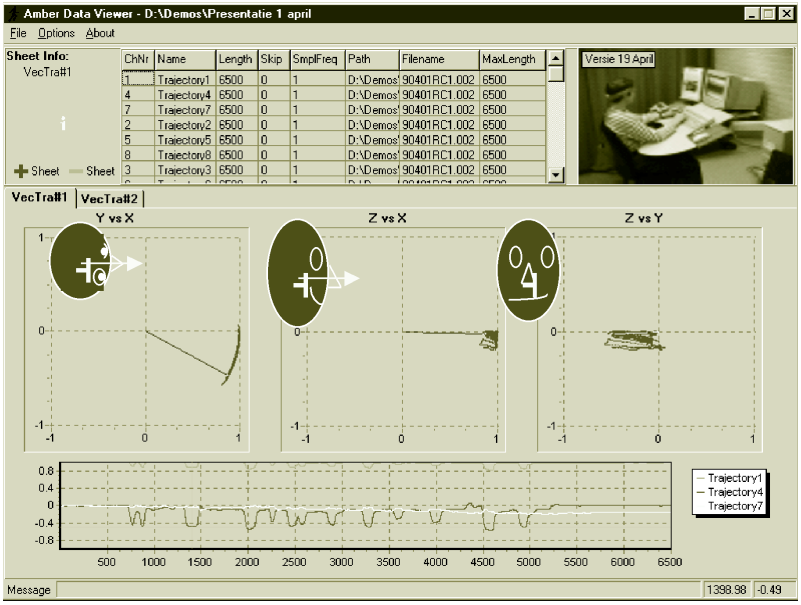


Figure 2
A typical example of a data (sub)set including the 3D orientation data in a time series applied to one of the three axes of the head coordinate system (bottom) also visualized in three projection planes or views (middle: top view, side view and frontal view).

[Adams, 1999; Burdorf, 1992; Burdorf, 1998; Burdorf, 1997; Burdorf, 1999; Engels, 1998; Fathallah, 1998; Frymoyer, 1983; Hoogendoorn, 1915; Keyserling, 2000; Mathiassen, 1993; Pope, 1989; Valat, 1997] and several data mining attempts are done with methods that supply highly mechanically incomplete or inaccurate data. This because of the unavailability of accurate methods to evaluate mechanical data in the field and the physical impossibilities (non-portability, limited field of view, high labour intensity) to apply more or less accurate laboratory based methods in the field. In only one study an attempt at directing data mining onto biomechanical risk predictors for work related complaints was done [Marras, 1995].

In this in vivo study the risk contribution of some 3D trunk kinematics[3] for low back disorders was assessed during occupational lifting. The study included workers from 400 industrial lifting jobs in 40 industries. From the analysis the most important risk factors could be identified. It was concluded that lifting frequency, load moment, trunk lateral velocity, trunk twisting velocity, and trunk sagittal angle were factors significantly increasing the risk of low back disorders. A model was derived that enables the assessment of risk of injury based on the characteristics of the lifting tasks. The developers went as far as realizing a commercial version including the risk evaluation model based on the data from the data mining study. Although much was learned about risk contributions, widespread application did not occur. Probable causes are practical reasons, e.g. the size of the system (it requires an exo-skeleton), and the requirement to manually assess weight and size of every load handled.

3 Movement ranges and paths.

Future challenges

With the future availability of small, truly ambulatory methods, like the one proposed in the introduction of this paper, for the first time large scale field experiments are possible, without compromises in mechanical accuracy, freedom of place and duration of the task or kinematic completeness.

New large scale studies are envisioned with ambulatory methods in which, for example, large numbers of subjects will be monitored during a representative period of their work, and in which subsequently the development of work related complaints is monitored in cohort type studies over 6 months to several years, all of this is very much in line with the study mentioned in the last paragraph. Then predictive relationships between aspects of the ambulatory assessed data and the risk epidemiology data are sought by applying data mining techniques. Hopefully the models obtained will deliver practical predictions for work related health risks, but will also contribute to the knowledge of the etiology of work related complaints.

Possible future consequences of all this could be that ergonomists will get a more quantitative basis for the ergonomically sounder design of work places and home furniture as well for ergonomically better organization of work and behaviour at work available.

They will also have a new generation of practical and affordable instruments available to assess working conditions where, in current ergonomical and rehabilitation practice, often no more than a 'gut feeling' is available.

References

- Adams, M.A., A.F. Mannion, P. Dolan. (1999). Personal Risk Factors for First-Time Low Back Pain. Spine **24** (23):2497-505
- Baten (a), C.T.M., J.H. van Dieën, H.J. Hermens, P.H. Veltink. Validation in Asymmetric Lifting
- Baten (b), C.T.M., J.H. van Dieën, H.J. Hermens, P.H. Veltink. Validation in Symmetric Lifting
- Baten, C.T.M., H. van Moerkerk, I. Kingma. (et al.). Net Sagittal Moment Estimation in Lifting through Application of Inertial Sensing. Clin Biomech. Submitted.
- Baten, C.T.M., H.J. Hamberg, P.H. Veltink. (1995). SAIBLE: A System for Functional Low Back Load Evaluation in the Field Combining EMG and Movement Sensor Data Using an Artificial Neural Network for System Calibration. Institut de Recherche en Santé et en Securité du Travail du Quebec, Montreal, Quebec, Canada. pp250-252
- Baten, C.T.M., H.J. Hamberg, P.H. Veltink, H.J. Hermens. (1995). Calibration of Low Back Load Estimation through Surface EMG Signals with the Use of Artificial Neural Network Technology

- Baten, C.T.M., P. Oosterhoff, I. Kingma, P.H. Veltink, H.J. Hermens. (1996). Inertial Sensing in Ambulatory Back Load Estimation
- Baten, C.T.M., J.H. van Dieën, H.J. Hermens, P.H. Veltink. (2000). Ambulatory Low Back Load Estimation
- Burdorf, A. (1992). Exposure Assessment of Risk Factors for Disorders of the Back in Occupational Epidemiology. Scand J Work Environ Health **18** (1):1-9
- Burdorf, A., G. Sorock. Positive and Negative Evidence of Risk Factors for Back Disorders. (1997). Scand J Work Environ Health **23** (4):243-56
- Burdorf, A., B. Naaktgeboren, W. Post. (1998). Prognostic Factors for Musculoskeletal Sickness Absence and Return to Work among Welders and Metal Workers. Occup Environ Med **55** (7):490-5
- Burdorf, A., A. van der Beek. (1999). Exposure Assessment Strategies for Work-Related Risk Factors for Musculoskeletal Disorders. Scand J Work Environ Health **25** (Suppl 4):25-30
- Engels, J.A., A.J. van der Beek, J.W. van der Gulden. (1998). A LISREL Analysis of Work-Related Risk Factors and Health Complaints in the Nursing Profession. Int Arch Occup Environ Health **71** (8):537-42
- Fathallah, F.A., W.S. Marras, M. Parnianpour. (1998). The Role of Complex, Simultaneous Trunk Motions in the Risk of Occupation-Related Low Back Disorders. Spine **23** (9):1035-42
- Frymoyer, J.W., M.H. Pope, J.H. Clements, D.G. Wilder, B. MacPherson, T. Ashikaga. (1983). Risk Factors in Low-Back Pain. An Epidemiological Survey. J Bone Joint Surg Am **65** (2):213-8
- Hoogendoorn, W.E., M.N. van Poppel, P.M. Bongers, B.W. Koes, L.M. Bouter. (1915). Systematic Review of Psychosocial Factors at Work and Private Life as Risk Factors for Back Pain. Spine **25** (16):2114-25
- Keyserling, W.M. (2000). Workplace Risk Factors and Occupational Musculoskeletal Disorders, Part 1: A Review of Biomechanical and Psychophysical Research on Risk Factors Associated with Low-Back Pain. Aihaj **61** (1):39-50
- Marras, W.S., S.A. Lavender, S.E. Leurgans. (et al.). (1995). Biomechanical Risk Factors for Occupationally Related Low Back Disorders. Ergonomics **38** (2):377-410
- Mathiassen, S.E., J. Winkel, K. Sahlin, E. Melin. (1993). Biochemical Indicators of Hazardous Shoulder-Neck Loads in Light Industry. J Occup Med **35** (4):404-7
- Pope, M.H. (1989). Risk Indicators in Low Back Pain. Ann Med **21** (5):387-92
- Valat, J.P., P. Goupille, V. Vedere. (1997). Low Back Pain: Risk Factors for Chronicity. Rev Rhum Engl **64** (3):189-94

Crime analysis on residential burglary data

Research by Manuel J.J. Lopez[2] , article by Jochen van der Wal[3]

Situation and problem

Burglary is an important reason for feelings of insecurity, in spite of a decreasing number of burglaries in the Netherlands over the past few years. With an estimated economic damage of 1,3 billion Euro, burglary is a serious problem, apart from the emotional consequences of residential burglary for the victims. This case describes the possible uses of data mining to create models for a better understanding of the process of residential burglary. The main goal is a better chance of catching the criminal and protecting homes from residential burglary.

Characteristics of the available data

In this case two different sources of information are used. The first contains information gathered by the police officer investigating the burglary. The second source contains demographical data, owned by the Dutch Central Bureau for Statistics (CBS).

As a part of the project to take a different view of the analysis of burglaries , an information model was constructed. The resulting information model contains all relevant data items needed for the data mining task to perform. The information model consists of the following items:

Offender data:

- Personal characteristics
- Social-economic background
- Burglary preparation activity
- Procedure of intrusion
- Behavior at the place of offence
- Behavior after leaving the place of offence
- Usage of burglary tools
- Preference for particular kinds of booty.

Victim:

- Personal characteristics
- Social-economic background
- Lifestyle/behavior
- Goods available
- Repeat victimization.

2 Dutch Police Institute (NPI), The Hague, The Netherlands, http://www.politie.nl/npi/

3 Korps Landelijke Politiediensten (KLPD), Driebergen-Rijsenburg, The Netherlands

Situation:

- Municipality information
- District and neighborhood information
- Premises information.

The largest part of the list above is available from the combination of the police and CBS sources. Unfortunately, in the test case some of the data items from the model above were missing. For the first try-outs only the available information could be used to perform the data mining task. In a follow-up project a team will revisit the victims of a burglary to get the complete data required by the constructed information model.

Choice of techniques and motivation

So far, only standard statistical techniques (frequency tables and Chi-square-check) are used to analyze the available data on burglaries.
To learn more about burglary, information about the offender, the victim and the situation have to be linked together. The connection between these three items can help to trace the burglar and protect citizens from residential burglary. The three items have a strong influence on each other. The (lifestyle of the) victim has an attraction for the offender, the victim changes the situation to protect his home, some kind of housings attracts some kinds of occupants, and so on. It goes beyond the scope of this case to explain all the details and backgrounds of the data mining techniques used. At this point it is sufficient to summarize these techniques. In this case, three different fields of data mining tasks can be studied.

Market based analysis

What are the patterns and combinations between the stolen goods? Can a link be detected between the manner of intrusion, stolen goods, tools used and possible behavior at the place of offence? To solve these questions, techniques like GRI[4] and factor analysis can be used. In our case, factor analysis is used. To detect possible connections between the theft of goods (e.g. taking the carkeys with the intention of stealing the car) the analysis is extended with the CHAID-analysis (Chi-square Automatic Interaction Detector).

Profile analysis

The goal is to analyze the situation of the burglary against a social-demographic background, to extract profiles of victims and to extract profiles of offenders from the available data. A combination of cluster analysis, compare means and discriminant analysis is used (see Sections 6.2.6, Clustering and 6.2.3, Discriminant analysis).

4 GRI stands for 'Generic Rule Induction', and falls into the category of associative techniques, often used in computerized knowledge systems.

Risk analysis

What is the burglary risk of the neighborhood or the risk that a person becomes a victim or offender? Besides the typologies from the profile analysis, situational information can also be used. CHAID is used for segmentation purposes.

Expected results

The police experiment to analyze residential burglary, resulted in deeper insights in:

- social-demographic situation typologies on district-level;
- social-demographic risk profiles on district-level;
- person related offender typologies;
- combinations between person characteristics, modi operandi and booty;
- the link between the stolen goods and the way of penetrating the home;
- home-jacking and the use of fire arms in residential burglary.

However, the available data is not sufficient for the stated purpose. After this project, an attempt will be made to gather the missing data, to satisfy the complete information model.
A sequel to this project is planned, to gather the missing data. The expected results are a deeper insight in the principles of burglary and the planning of more effective police surveillance on basis of the detected 'hot spots'. The police can locate the homes with a higher burglary risk and give the occupants information on how to protect their homes. The detected profiles of the offenders give the police investigators better starting-points to solve the case. In general, it is expected that data mining will give the police a better basis for preventing and solving burglaries.
The general goal is to use the available police manpower in a smarter way, by means of extracting more information from the available data.

Future prospects

We don't know what the future will bring, but the results from the data mining activities are promising. Data mining on the available data may, in the future, help the police to find 'attractive' residences. This could result in special attention to the residences with a higher risk. A prevention team would provide the occupants with advice on preventing residential burglary. The extracted profiles indicate the vulnerable groups of inhabitants who should be considered for extra attention and a special treatment.

The ultimate future perspective is that a police officer investigating a burglary would store all the information gathered in his handheld computer. The computer would return a description of the most likely offender and the place of the house he probably will break into next. The police officer will be guided more

directly and more efficiently. The officer will be directed to a certain place, on a certain time to look for somebody with well-defined characteristics. There is no place for wasting time with vague directives, posting somewhere with a lack of information, or not knowing what to look for anyway.
Another promising technique is the calculation of the offender's residence, from his geographic home breaking pattern. Research in this field done in Canada could probably be adapted or translated to other situations.

3.2.5 Prediction

One of the more complex tasks, prediction often involves deriving patterns from a training set, thus building a model of the population's behavior that incorporates predictor variables and dependable variables. The model is then fed with new predictor data to produce estimations of the dependable variables.
In the strict sense of the word predicting would involve a time difference between the predictor variables (now or in the past) and the dependable variables (in future). However, in statistics or data mining 'predicting' is also often used for the estimation of unknown dependable variables at present or in the past. Our first example is based on the known responses of a test set of customers and tries to predict the set of prospects who are most likely to buy a caravan insurance policy. The next example is of a more complex nature, the authors investigate the existence of predictable sequences of events in the financial markets.

Analytical customer relationship management for insurance policy prospects

Peter van der Putten[1]

In marketing there are two opposed approaches to communication: mass marketing and direct marketing. In mass marketing, a single communication message is broadcast to all potential customers through media such as print, radio or television. Such an approach implies high wastage: only a small proportion of the prospects will actually buy the product. As competition increases and markets get more fragmented the problem of waste worsens. Moreover, in spite of huge investments in market research and media planning, it is still hard to quantify the benefits of mass marketing.

These developments have led to an increased popularity of direct marketing, especially in the sectors of finance, insurance and telecommunication. The ultimate goal of direct marketing is cost-effective, two-way, one-to-one communication with individual prospects. For this it is essential to learn present and predict future customer preferences. In today's chaotic business environment, customer preferences change dynamically and are too complex to be derived straightforwardly.

Continuous mining of customer behavior patterns may offer a flexible solution to this problem [Putten, 1999]. The classical application of data mining for direct marketing is response modeling. Usually, the relative number of customers that

1 Drs P. van der Putten, pvdputten@hotmail.com. pvdputten@liacs.nl, Leiden Institute of Advanced Computer Science, Leiden University, Leiden, The Netherlands

respond to direct mail is very low (5% or less). Predictive models can be built to identify the prospects most likely to respond. Historical data on previous mailings or product ownership are used to construct the model. The resulting model can be applied to filter prospects from the existing customer base or from address lists acquired from list brokers. We present a case from insurance marketing to illustrate this approach.

Business problem

The business objective in the insurance case was to expand the market for an existing consumer product, a caravan insurance, with only moderate cost investment. Data mining analysis should answer the following question: can we predict who would be interested in buying a caravan insurance policy and explain why?

Actually, this real world business case was re-used for the CoIL Challenge 2000, a data mining competition organized by CoIL, the European Network of Excellence for Computational Intelligence and Learning [Putten, 2000] [2]. The problem description and data were posted on the web and participants had a little more than one month to send in results. In the remainder of this article we will focus mainly on the challenge results for the predictive data mining task.

Data mining solutions

The data used in the case were very similar to data in other direct marketing projects for other financial clients. Each customer was characterized by a selection of 95 attributes. The attributes could be divided in two groups. The product usage attributes defined the product portfolio of an individual customer, consequently these attributes can be considered internal (company owned), behavioral attributes. We also purchased sociodemographic survey data that had been collected on zip code level. All customers belonging to the same zip code area have the same value for these attributes. This data included information on education, religion, marital status, profession, social class, house ownership and income. Most participants performed extensive data preparation to get the best input for the models, including using derived attributes, recoding attributes and selecting attributes.

To select prospects a model had to be constructed to predict the attribute 'owns a caravan policy', given all other attributes. For the challenge a random sample, the training set, was drawn from the customer base. This set was used to construct the models. The training set contains over 5,000 descriptions of customers, including the information whether or not they have a caravan insurance policy. Furthermore, there was a test set containing 4,000 customers from whom only the organizers knew whether they had a caravan insurance policy or

2 On the 'The Insurance Company Benchmark Homepage' the problem statement, data and results are made available for research purposes (http://www.liacs.nl/~putten/library/cc2000/).

not. The test set was used to determine how well the model performed for cases that were not used in training.

For a prediction task like this, the underlying problem is to find the subset of customers with a probability of having a caravan insurance policy higher than some boundary probability. The known policyholders can then be removed and the rest receive a mailing. The boundary depends on the costs and benefits such as the costs of mailing and benefit of selling insurance policies. In the case of the challenge, we simplified the problem: we wanted the participants to find the set of 800 customers in the test set of 4,000 customers that contained the most caravan policy owners. For each solution submitted, the number of actual policyholders was counted and this gave the score of a solution. So in summary, the prediction model had to be able to calculate a reasonable estimate of the probability that a customer that was not in the training set owned a caravan policy.

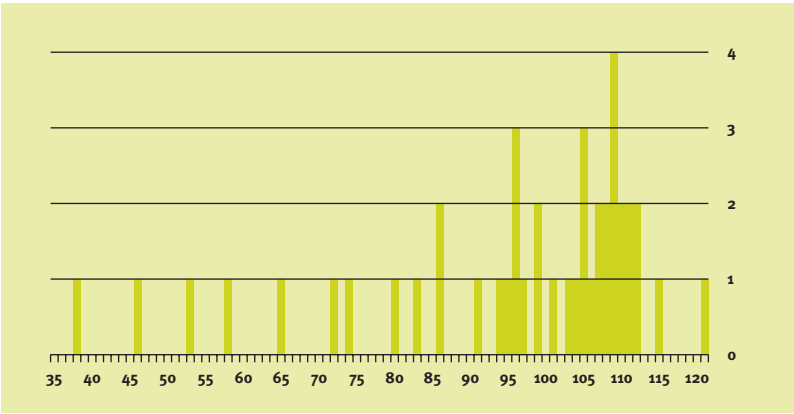


Figure 1
Prediction results. Frequency distribution of the number of real caravan policy owners in the submitted selections. Random selection would result in 42 policy owners, so in most cases there was a substantial improvement in response rate by using the prediction model.

In the end 43 solutions were sent in. In the majority of the cases approaches from more than one area in computational intelligence and statistics were used. The frequency distribution of scores for the prediction task are displayed in Figure 1. The maximum number of policy owners that could be found was 238, the winning model selected 121 policy owners. Random selection results in 42 policy owners. Our standard benchmark tests result in 94 (k-nearest neighbor), 102 (naïve Bayes), 105 (neural networks) and 118 (linear!) policy owners. A wide variety of methodological approaches were used by the participants including boosting, bootstrapping and cost-sensitive classification. Algorithms used included standard statistics, neural networks, evolutionary algorithms, genetic programming, fuzzy classifiers, decision and regression trees, support vector machines, inductive logic programming and others (see Part 6, Data mining methods and technology). A general result reported by most participants was that the product usage variables were better predictors than the sociodemographic variables. This was to be expected. Previous customer behavior is the

best predictor for future behavior. Furthermore, the basic assumption behind the sociodemographic variables, that every one in the same zip code area is similar, is often violated in practice. However, these zip code-based variables can still be valuable, for instance for descriptive data mining or when product usage is not available.

The challenge confirmed our findings in the prior commercial project: that by using data mining prediction a clever selection of prospects could be made. Such models can be used in situations where manual selections are not feasible. For example, marketeers might be able to make a coarse grained manual selection by using their marketing knowledge. Models can then be used to make a fine grain selection from this set. Furthermore, marketeers sometimes lack the time to experiment with different queries to determine an optimal mail shot. Automated modeling can help in this case as well. In both cases it is important that the model can be explained and that the prediction performance on new cases is demonstrated.

A number of findings in this direct marketing case might be relevant for other applications.
First, the spread in the prediction scores is rather large. Apparently, expertise with the method is required. To let data mining grow into a tool for end users, this expertise must be made explicit, formalized and automated where possible. This should include steps that go beyond the core algorithm, such as data preparation (e.g. feature selection) and evaluation (application specific evaluation measures, boosting, combining models).
Second, an approach which was suggested to be the most prudent in the after challenge discussions was to 'try the simplest first and be self-confident'. On real world prediction problems like the one in the challenge, one should try a wide variety of approaches, starting with the simplest ones, because they seem to work best. This indicates that simple computational learning algorithms can play an important role, when used as alternatives to standard statistical techniques and thus improving the choice range of algorithms. On the other hand, simple statistics should always be included.

Vision for the future

With respect to current marketing applications, there are a lot of similar useful predictions that can be made, such as predicting customer retention or estimating turnover potential.

Looking into the future always boils down to some degree of speculation, but it is reasonable to assume that the trend towards more personalized and ultimately one-to-one marketing will pull through. Most modern marketeers claim

that this is the holy grail of marketing, however, it will most certainly make life a lot harder for them. As public awareness on privacy issues increases, companies that perform marketing on a non-selective, wasteful manner will simply be neglected by customers. A so called one-to-one relationship must be meaningful from the perspective of the customer too. Privacy protectors and data mining might form a paradoxical alliance with the same objective: less undesirable sales contacts.

Because of this trend towards personalization, the average mailing campaign size will be orders of magnitude smaller, so the selection quality requirements will increase. On the other hand, the amount of campaigns might grow dramatically. So manual selection or even 'manual' construction of data mining models will become merely impossible. In modern analytical customer relationship management and marketing campaign management software this trend towards automation of the selection process is clearly visible.

The next step would be that instead of off-line, centrally organized campaigns marketing actions will be real-time, distributed and customer driven. For instance, any change in customer profile data might result in a changed product propensity level and fire off a marketing action. Even changes in customer profiles of people that resemble a customer might result in such an action.

Finally, it is reasonable to expect that customers will start to use mature, automated mining software to turn the tables. Intelligent agents will scour electronic markets to search for necessary, interesting and useful products. Customers will keep profile information private, and only release anonymous profile information, if they directly profit from it.

If this black-and-white scenario becomes reality, data mining will be a blessing for consumers, but for marketeers it will seem to be more like a devil in disguise. But in the end a more efficient exploitation of marketing budgets will benefit both.

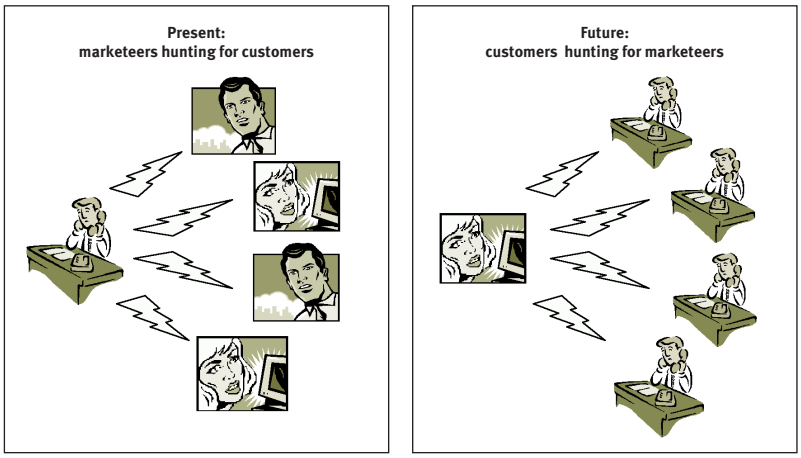


Figure 2
In future customers will use data mining to find the best product.

References

- Putten, P. van der. (1999). Data Mining in Direct Marketing Databases. In: Walter Baets. (ed.). Complexity and Management: A Collection of Essays. World Scientific Publishers, Singapore
- Putten, P. van der, M. van Someren. (eds.). (2000). CoIL Challenge 2000: The Insurance Company Case. Sentient Machine Research, Amsterdam. Also a Leiden Institute of Advanced Computer Science Technical Report 2000-9. June 22

Pockets of predictability in financial markets

Willem-Max van den Bergh[3], *Jan van den Berg*[4]

Predictability of financial market returns

In this chapter we address the potential of advanced data mining techniques in financial time series analysis. But first of all we must deal with the question of how return series[5] can be more than accidental strings of random, unrelated market-level changes. Actually, we bring the efficient market hypothesis (EMH), which states that current market prices reflect all relevant information, under discussion. By definition, new information arrives randomly and thus, as many finance scholars will agree, a securities market that does not exhibit a random walk cannot be efficient. On the other hand there is no doubt among practitioners that financial market returns are predictable. However, most admit that this predictability varies over time. Traders make money by exploiting market opportunities. By doing so they have become very good at recognizing momentary (local) features of the market. This might explain why technical analysis is so popular among traders. Technical analysis gives a description of the market that is very different from the usual statistical description. Technical patterns are truly local and only once in a while, a distinct buy or sell signal emerges. The focus of traders on tools that give a local description of the market cannot be ignored. These situations, also called local patterns or pockets of predictability, are hard to find with conventional techniques that try to model the overall market structure. In the last decade, the notion that there are periods of higher than normal predictability in the market has also entered financial literature, see for an example [Pesaran, 1995] and [Dunis, 1996]. Since these pockets of predictability are very specific and undoubtedly different from the normal everyday market behavior, we will refer to them as exceptions.

Financial markets are characterized by a large number of participants, each having a different appetite for risk, a different time horizon and different motivations and reactions to unexpected news. In the circumstances it would come as

3 Drs W.-M. van den Bergh, vandenbergh@few.eur.nl, Erasmus University Rotterdam, Department of Financial Economics, Rotterdam, The Netherlands

4 Dr Ir J. van den Berg, jvandenberg@few.eur.nl, Erasmus University Rotterdam, Department of Informatics and Economics, Rotterdam, The Netherlands, http://www.eur.nl/few/people/vandenberg

5 Return series relate here to the relative value change (in %) of stocks between time t-1 and t.

$$return = \frac{stock\ value_t - stock\ value_{t-1}}{stock\ value_t} * 100\%$$

a surprise, if all these complex interactions were to average out in a linear fashion. Consequently, new mathematical and statistical tools, which rely heavily on the analysis of non-linearities, are being developed. Many of these new tools can be characterized as data mining techniques, i.e. methods of exploratory analysis looking for meaningful patterns and rules. Some of them use a bottom-up approach, called knowledge discovery, where no prior assumptions are made; the data is allowed to speak for itself. Undirected knowledge discovery [Berry, 1997], a main issue in this article, attempts to find patterns or similarities among groups of records without the use of a particular data field or predefined classes.

A quick search in the existing literature of undirected knowledge discovery did teach us that in general, only modest attention is put on the discovery or learning of exceptions. Instead, most algorithms put the emphasis upon the discovery of the common rules. Exceptions are often an afterthought. In this chapter, we explicitly look for exceptional patterns having certain predicting power. The remainder of this chapter presents two main issues. The first is more economically oriented and pursues theoretical backgrounds for the existence of exceptions in financial markets. We will go into some very specific exceptional situations in financial markets like stock market crashes. Some novel attempts to model market behavior during these situations are presented in order to get more insight into the nature of such exceptions. Our main focus will be on the system dynamics of financial markets and we will look at different fuzzy states (market regimes) and the transition between these. Analogies with physical systems and the existence of exceptions like avalanches and traffic jams will be discussed in an intuitive manner. In the second section a Competitive Exception Learning Algorithm (CELA, see below) is introduced where exceptional patterns are inferred from a given set of data pairs. Though our inspiration for developing CELA stems from the wish to analyze financial time series and finding pockets of predictability, we think the algorithm is applicable in a much broader field.

Exceptions in financial markets

According to many advocators of the efficient market hypothesis, phenomena like bulls[6], bears[7] and market bubbles are viewed as accidental strings of random, unrelated market-level changes. But how irrational are bubbles really? Recent academic literature [Treynor, 1998] indicates that it is not obvious that, if the market level is temporarily perturbed, equilibrium forces will return it to its original level. If, for example, the perturbation drives the market level up slightly, the bulls will gain at the expense of the bears. Because of the wealth shift, the market will accord greater weight to the bulls than formerly and less weight to the bears. So, now the equilibrium market level has risen. Although the actual price change may have been a complete surprise, the consecutive shift to a

6 Bull: buying stocks to cause price rise, or in expectation of a price rise.

7 Bear: selling stocks to cause price drop,planning to buy back larger quantities at a lower price.

new equilibrium may be quite deterministic. Obviously a potential regime shift is determined by the opportunities bears have to continue their game: which options do they have under these extreme conditions? If the perturbation is quite large, and many investors reach the end of their financial capacity, the new equilibrium may be some distance from the old. Furthermore, the transition will not be smooth and shorter or longer periods of overshooting may exist, which may be related to the way the market is organized and is able to process the order flow during hectic periods. The information available to investors about what is actually happening will certainly influence market behavior. We will call all such deviations from random walk exceptional. It is not necessary to argue that early detection of these exceptions is of great value to investors and risk managers.

The nature of market crashes

One of the most interesting examples of exceptional situations in financial markets is a market crash. Such crises show some notable similarities with traffic jams. The transaction flow in financial markets knows the same sudden transition between orderly normal behavior and sudden explosive and seemingly irrational bursts, eventually followed by one or more collisions, when market parties default. For both, financial crises and traffic jams, it is questionable whether they are predictable. Undoubtedly this is a hot topic: J.P. Morgan has an 'Event Risk Indicator'[8], Credit Suisse First Boston an 'Emerging Markets Risk Indicator', and Lehman Brothers a 'Currency Jump Probability Measure'. However, if we may believe The Economist ('The Perils of Prediction', August 1999): "Don't expect them to work ... sophisticated stuff, to be sure. Yet the real question is whether this improves on what investors use today...!" On the other hand, for complex phenomena like the weather there is a clear breakthrough to better insights and new technological tools.

Different market regimes as fuzzy system states

The understanding of traffic jams increases little by little, at the same time providing insights for financial markets. Most models used originated more or less in theoretical physics. Consequently, a first source for these models is professional literature like 'Nature, Science and Physica'. E.g. recent research at Daimler Benz [Kerner, 1977] emphasizes the relation between the development of traffic congestion and crystallization. A distinguished feature of this process is very abrupt state transition. A single dust particle can act as a catalyst in a supersaturated solution and lead to sudden crystal formation. The new crystals act themselves as catalyst and crystallization spreads like a wave. Thus, the process leading to traffic congestion shows several distinct states. When flow increases, we observe 'clusters' of high traffic density moving with increasing intensity backwards. Above a distinct density all car speeds become, so to

8 http://www.jpmorgan.com/CorpInfo/PressReleases/1998/01301998_ERI.html

speak, 'synchronized'. One single speed is imposed on all cars and overtaking becomes difficult. When density increases even further, collisions may occur.

In financial analogy, one single sales order might well trigger many others and lead to a high volatility. And when certain boundaries are encountered, a crash might even follow. When volatility increases, price rises and price drops of increasing magnitude follow each other even faster and suddenly the price breaks through in a distinct direction. Generally this takes the shape of a persistent price drop, but sometimes we observe a sharply increased price level (usually referred to as euphoria, but holders of a short position will definitely disagree). It seems as if returns freeze for some time at a more or less constant value. This resemblance to state-change in pure physics suggests the investigation of the properties of physical models for predicting market behavior. Identification of different states and the exact context for state transitions is crucial in these models. However, the financial equivalent of such models is truly complex, mainly on account of the many possible ways open to individuals to react to sudden market movements. Furthermore, state transitions will undoubtedly be fuzzy and less distinct than in many physical situations ('ice below zero and steam above 100° Celsius'). Advanced data mining techniques used on databases containing the full contextual market setting of large amounts of transactions have a high potential in this respect, especially if they are able to deal appropriately with fuzzy events (see also Section 6.2.16, Fuzzy logic techniques).

State transition and options theory

Financial theory and physics are less dissimilar than one might think. The kernel of the options pricing model as formulated by [Black and Scholes, 1973] is a differential equation that is not different in any respect from what is known in physics as 'heat transfer equation'. If a source of heat is put into contact with some fluid, the temperature of the fluid will rise gradually until the boiling-point is reached. The analogy in options pricing: if the price of the underlying value of an option changes, the pay-off for the owner changes gradually until the strike price is reached. Beyond that point the pay-off is fixed at the exercise price. Clearly the relation between changes in the underlying value and the pay-off of an option is non-linear. This situation is quite customary in financial contracting. If a fixed interest bond is used by a firm to finance an asset position with a variable value, the pay-off to the bondholder is constant (principal plus interest payments) if, at maturity, the asset is worth more. If the asset is worth less, the bondholder only gets this lower value, i.e. what remains in case of default.

The higher the threat of default, the more it will influence the behavior of the contractual parties, which in return may influence market transactions. In other

words, in the proximity of position limits, the return/volatility pattern of asset prices may differ from the usual, normal pattern. Things will become really difficult, if there is no counter-party for a transaction or if the maximal order processing capacity of the market has been reached. The only way open may be a stop-loss order resulting in a reinforcement of the current market direction. Others are triggered to react and so on. Clustering increases and in extreme conditions crystallization starts and the market freezes and the returns maintain to fall (or rise) for some time. Such events should be identifiable early in market return time series by shorter or longer periods of serial correlation. Advanced data mining techniques appear to have great potential for finding patterns in the market context, when distinct behavior has occurred. The hunt should be for a method of separating these specific (and sparse) contexts from situations that are not followed by abnormal market behavior.

From a physically orientated explanation for market crashes we can formulate the following hypotheses:

- Crashes of all magnitude are possible, only limited by the amount of all positions taken.
- The probability for a crash may be higher than predicted by standard risk management models[9] based on standard deviation or semi-variance.
- There is no necessary relation between large returns and special news events[10].
- There is no typical time interval between two crashes.

Institutional and behavioral bias

The probability of crashes is, apart from the order flow itself, determined by 'institutional' properties like the way the market is organized, order processing, availability of information about recent prices and order positions. As mentioned earlier, the way positions are limited and other regulations like 'circuit breaking', i.e. stopping the order flow for a given period when the market gets over-heated, will certainly play an important role. Market organization is important for the speed of order processing in active moments [Martens, 1998]. Capacity and reliability of the existing automated system are obviously important.

A large and rapidly growing body of literature attributes various stock market anomalies to behavioral biases. An example is the finding of [De BondtThaler, 1985] that people tend to overreact to dramatic events: e.g. investors seem to attach unrealistic probabilities to stock market crashes. Some researchers mention what is called 'representativeness bias', which, simply put, means that people tend to think: "if it walks like a duck and quacks like a duck, it must be a duck." [Scott, 1999]. In other words, if a given market context somehow resembles a situation of the past that was followed by a specific market behavior like

9 The Black and Scholes [Black and Scholes, 1973] model discussed earlier is one of these.

10 Note that this is incompatible with the (among econometricians) popular 'jump-diffusion model', which relates large external shocks to large market movements.

a crash, they get anxious and react accordingly. The same may occur, when people get euphoric when they recognize a former prosperous setting for a given stock or the whole market. In this sense the market reaction becomes a 'self-fulfilling prophecy'. In our view, such findings provide a strong argument for the use of advanced data mining techniques that map historic abnormal (exceptional) market behavior to the contextual environment immediately before and during such events.

Mining with a Competitive Exception Learning Algorithm (CELA)

In the previous section we discussed the typical behavior of financial time series in exceptional situations. We will now consider some properties of a truly intelligent agent (either a person or a computer program) being able to early detect such events. In mathematical terms, his (or its) task is to correctly predict a future market state $y(t+1)$ given m historical states $x(t)$, $x(t-1)$, $x(t-2)$, ..., $x(t-m+1)$. Here, any historical state $x(t-i)$ is supposed to potentially affect the future market state. E.g., by purely analyzing the time series $y(t)$, $x(t-i)$ may simply describe the historical market state $y(t-i)$ itself. Otherwise, $x(t-i)$ may be composed of 'environmental' state values at time $t-i$ (like economic news facts about rental, unemployment, growth rates, etc.) or institutional setting, which, when changing, have their impact on future market states. The Competitive Exception Learning Algorithm (CELA) introduced here is especially designed to play the above mentioned role of 'intelligent agent'.

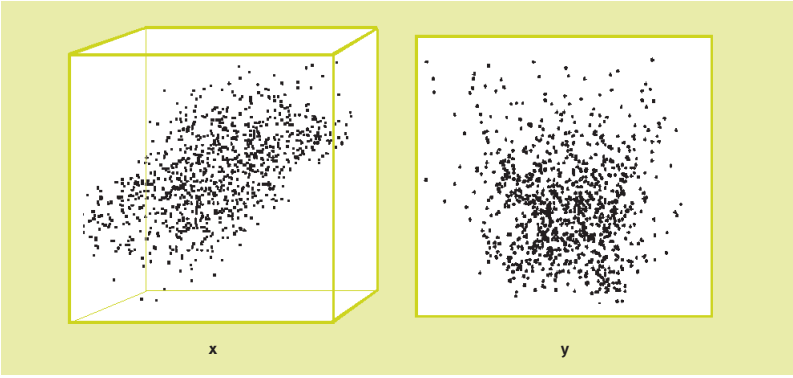


Figure 1
The 3-dimensional input X and the 2-dimensionial output data Y of the working example.

As a working example to illustrate the operation of CELA, we will use a simulated data set containing conditional serial dependencies, in line with market concepts set out previously. We assume that prices are bid up in a bullish market or bid down in a bearish market, resulting in shorter or longer periods of serial correlation in returns. So, a set of data pairs has been constructed having the above given characteristics of noise trading: output space *Y* (see Figure 1) consists of data points (cases) representing the future 1-period return and the future 5-period volatility. The input space *X* consists of data points representing

the standard deviation over the last 5 periods, the average return over 5 periods and the 1-period past return. Each new return is generated using a transformation from a uniformly distributed random variable to a normal distribution under 3 regimes (2 of which are exceptional in the sense that they generate serial dependency). The goal of CELA is to infer the regimes together with the corresponding exceptions.

Generally speaking, the task of the algorithm is to infer a conditional relationship between an M-dimensional input space (*X*) and an N-dimensional output space (*Y*). A more formal description of the algorithm is given in [Bergh, 2000]. Five distinct steps can be distinguished:

1 Determine clusters in *Y*, using a competitive learning approach, in such manner that overall fuzziness is minimized.
2 Determine the unconditional output fuzzy membership distribution (UOD) for these output clusters.
3 Determine clusters in *X* where the conditional output fuzzy membership distribution maximally deviates from the UOD as found in the previous step. In this step we also use a competitive learning approach.
4 Identify a fuzzy rule base from input and output clusters as found in the previous steps.
5 Perform a function approximation in order to estimate an output estimate for each individual input case.

We show the result of the first 4 steps on our working example in Figure 2. The number of clusters in our method is user defined. We chose 4 output clusters, 3 of which have their centroïds[11] at a fixed location in a corner of the output space (see the right-hand side in Figure 2). The location of the 4-th cluster centroïd is determined using the above mentioned competitive learning approach. The centroïds are located so that all cases are as near as possible to one of these centroïds. All cases compete, in a way, to attract a centroïd. The clustering is characterized by minimal fuzziness, since the average membership of each case for the cluster with the nearest centroïd (the winner cluster) is maximal. Each case is member of all (in the example 4) clusters, while the memberships value for the winner cluster is highest. The membership values sum up to 1. In our approach we consider cluster memberships as inversely related to the squared distance from the centroïds.

In the next step these individual memberships are averaged over all cases, arriving at the unconditional output distribution (UOD). The percentages in the square on the right-hand side represent this UOD, which equals $\bar{\mu}^{Y}$ = (0,0805, 0,7427, 0,0852, 0,0916). Thus, having no further information about a specific case, it's expected relative membership is 74,27% to the largest class, 9,16 % to the next, etc.

11 Center of gravity of a shape or point cloud.

The third step entails partitioning of the input space. This step is crucial but complicated. The input clusters have been situated so that the local behavior of the associated output is as different from the UOD as possible. In other words, the competitive algorithm searches for exceptional events that jointly occur with a specific input context. This time, the input cases x_p compete for cluster centroïds $\bar{x}_b$ (b indicating the cluster number). Cases that, with respect to their output, deviate strongly from the UOD (measured by the Euclidean distance between the local membership vector and the UOD) have a stronger attraction to input centroïds than 'normal' cases that are distributed more or less similarly as the UOD.

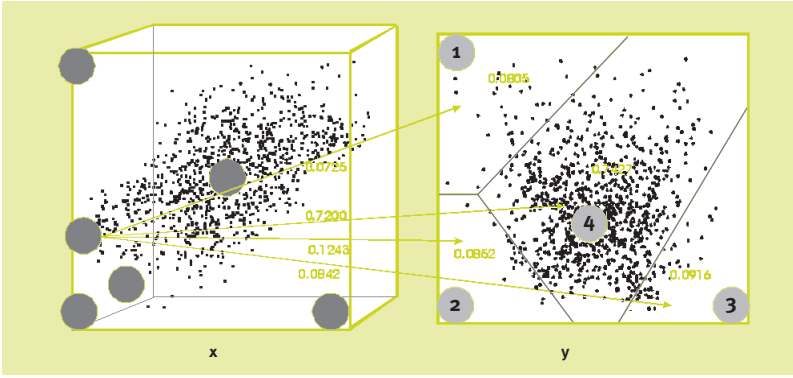


Figure 2
Partitioning of both the input and the output space after learning. The output space Y is divided into 4 fuzzy classes, each predicting a specific return development. For example 1= low at t, high at t+1, 2= low at t, low at t+1, 3= high at t, low at t+1, 4 = around average. Exceptional events at t reflect in other than average memberships for clusters in the input space. leading to a different membership probability for the clusters in the output space of expected returns.

Output values y_p deviating from the UOD are considered 'potentially exceptional' and thought to emit an output exception signal (OES), based on the Euclidean norm as mentioned earlier. Only if they systematically correspond to specific locations in the input space, however, are they actually considered exception or exceptional pattern. This step of the CELA algorithm uses the joint product of OES and winner input cluster membership and seeks to minimize its average value over all cases. The resulting input clustering is characterized by minimal fuzziness as was the output clustering, but now with respect to the characterization of exceptional patterns. Once more, each case is member of all (in the example 6) input clusters.

In step 4 of the algorithm we calculate the weighted average output membership distribution per input cluster. The input memberships are used as weights. The intuition behind this step is clear: the more a specific input is member of a given input cluster, the more we expect its output behavior to behave like the average output behavior of that cluster. This average cluster behavior (a 'local' output distribution) is thus conditional on the input cluster and can be conceived as being located in the centroïd. In this sense such local output distribution can be seen as a rule: knowing we are in a given input cluster centroïd, we have a specific output distribution which differs from its unconditional equiva-

lent. All rules, i.e. the local behavior of all cluster centroïds apply to a certain extent to each input case, depending on the membership for the input clusters. Hence, we may typify the rules as fuzzy. Consequently, input locations lying far away from any cluster centroïd will hardly follow any rule and cannot be classified as exceptional.

Once the fuzzy rule base has been identified, we are also able to estimate (step 5) an output location for any input data point. First we calculate the relative membership for each of the input clusters and next we derive the conditional relative output cluster membership based on the fuzzy rule base and using the input cluster memberships as weights. From this estimated output membership distribution we calculate the expected value as a weighted average from the cluster centroïds.

The 3-dimensional space at the left of Figure 2 shows the result of the input clustering. For one of the input clusters the conditional rule as determined on the working example is shown as an example. We notice that the conditional distribution $\bar{\mu}^{Y} \mid \bar{\mu}_{c}^{X}$ = (0,0725, 0,7200, 0,1243, 0,0842) which clearly deviates from the UOD of (0,0805, 0,7427, 0,0852, 0,0916). Notably, the conditional membership of 0,1243 for the third cluster is higher. As can be seen from the figure, this cluster has a relatively low return and a relatively low volatility. In economic terminology, we may characterize it as 'a price drop'. The input cluster has a 5-day volatility that is low to normal, a low 5-day return and a relatively low 1-day return. An economic description of this situation might be something like: 'market is falling'. Now we can express the earlier rule as: 'If the market is falling, there is a higher than normal probability of another price drop'. This rule, although it may look trivial, exactly mimics an element of the simulated behavior of the time series during an exceptional regime. In this sense it is a good illustration of 'opening the black box' by studying the fuzzy rule base resulting from CELA.

Conclusions and prospects for the future

Intelligent data mining has the potential to learn very specific characteristics of financial market behavior. Theories that incorporate the institutional setting of the market as well as the behavioral aspects of all market participants and take a true local view of the conditional setting imply non-linear market behavior. In this contribution we have shown some extra capabilities of specialized data mining techniques which are able to deal with such underlying non-linear relationships. The Competitive Exception Learning Algorithm (CELA) is an example of such a technique with potential value for professional investors as well as regulators. It may help not only to anticipate irrational market movements, but also to understand them better. Of course, extensive future research is necessary.

We are convinced that new highly dedicated algorithms (like CELA) will be detected and eventually become available as standard, albeit specialized, tools. The possibility of such tools to get insight in the 'rules behind the system' makes them appropriate for embedding in highly domain-oriented, but user-friendly decision support systems. Without doubt, they will make financial markets more efficient in their important task of risk-diffusion between individuals.

References

- Bergh, W.M. van den, J. van den Berg. (2000). Competitive Exception Learning using Fuzzy Frequency Distributions. ERIM-report ERS-2000-06-LIS. http://www.erim.eur.nl/. Publications in Management, Section Business Processes, Logistics and Information Systems
- Berry, M.J.A., G. Linoff. (1997). Data Mining Techniques for Marketing, Sales, and Customer Support. John Wiley & Sons, New York
- Black, F., M. Scholes. (1973). The Pricing of Options and Corporate Liabilities. Journal of Political Economy **81**:637-654
- Bondt, W. de, R. Thaler. (1985). Does the Stock Market Overreact? Journal of Finance **60** (3):793-805
- Dunis, Ch. (ed.). (1996). Forecasting Financial Markets, Exchange Rates, Interest Rates and Asset Management. John Wiley & Sons
- Kerner, B.S., H. Rehborn. (1977). The Physics of Traffic Jams. Physical Review Letters. November 3
- Martens, M. (1998). Price Discovery in High and Low Volatiliy Periods: Open Outcry Versus Electronic Trading. Journal of International Financial Markets, Institutions and Money **8**:243-260
- Pesaran, M.H., A. Timmermann. (1995). Predictability of Stock Returns: Robustness and Economic Significance. Journal of Finance **4**:1201-1228
- Scott, J., M. Stupp, P. Xu. (1999). Behavioral Bias, Valuation and Active Management. Financial Analysts Journal **55** (4), July/August
- Treynor, J., Bulls, Bears and Market Bubbles. (1998). Financial Analysts Journal **54** (2):69-89

3.2.6 Matching

Peter van der Putten[1]

Matching supply and demand is a core process for industry and government, whether it concerns searching for products, services, persons or information. In this chapter we will give a quick introduction to the data mining technologies used for matching. Furthermore, we will illustrate this with two cases from the profit and non-profit sector, hunting for criminals and hunting for jobs. We will end with a vision of the future.

How does matching work?

Finding the best products to buy, selecting the best prospects to sell to, allocating the best matching resources — human or other — and serving the right information to citizens can all be seen as examples of matching processes. Given a demand, the best offers have to be found and evaluated (or the other way around). Typically, supply and demand do not correspond perfectly.

Classical information technologies, such as standard database queries, only offer a limited solution to this problem. If a user is not capable of formulating an exact, errorless query, a conventional query system is likely to return nonsense instead of a reasonable answer. Difficulties also arise when an exact query can be formulated, but no data satisfies all the specified criteria. The cause of these problems is that conventional search techniques rely on Boolean, all-or-nothing logic, which makes the search rigid and brittle: a fraction of an inch can make a world of difference. People are obviously able to handle fuzzy descriptions and match them to 'objects' in the real world. Conventional information technologies are hardly capable of doing this.

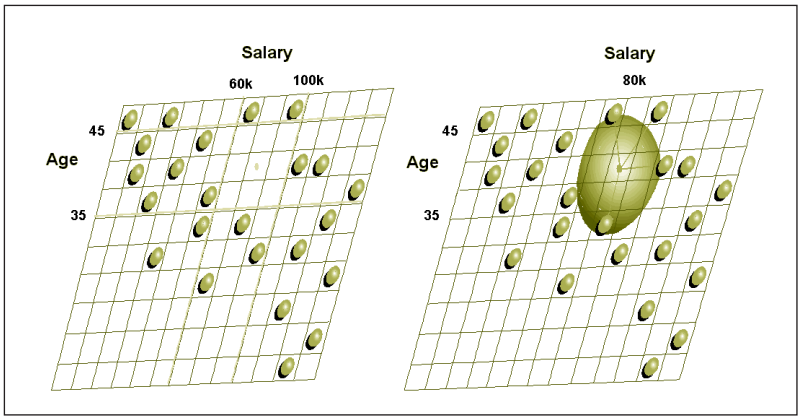
Nearest neighbor search

To support matching processes fully, alternatives should be found that match to a certain degree. Data mining offers a wide variety of technical solutions to support the matching process, including prediction and clustering. But the most basic and simple approach to this problem is nearest neighbor search (fuzzy matching).

The idea behind nearest neighbor matching can be explained with the following example from human resources matching (Figure 1). Assume we are looking for four persons that are around 40 years old with a salary of around 80,000 Euro. Each dot is a person in the database. If standard database queries are used (left), various strict search criteria should be set, for example: 35-45 years old and salary between 60,000 and 100,000 Euro. The user doesn't know before-

1 Drs P. van der Putten, pvdputten@hotmail.com, pvdputten@liacs.nl, Leiden Institute of Advanced Computer Science, Leiden University, Leiden, The Netherlands. This section was written, while he was a consultant for Sentient Machine Research.

Figure 1
Matching as a data mining task: nearest neighbor search



hand how many persons fit the criteria perfectly (none in this case) and must consequently tune the constraints over and over again to get a selection of reasonable size and quality.

Nearest neighbor search offers a solution (right). The user describes the ideal person to be found — 40 year old, 80,000 Euro salary. Then the nearest neighbor algorithm starts with selecting all persons that match perfectly (again none in this case). Next, all criteria are loosened simultaneously, until the requested amount of people is found. So a nearest neighbor match engine can be seen as a search engine like AltaVista, but then for structured database information. In reality this is often implemented by calculating the distance of the relevant properties for each record to the query point, in this case the 'ideal person'.

The nearest neighbor algorithm offers several additional information elements. The distance from a person (large dot) to the search profile (small dot) gives an indication of the relative match. A user can attach weights to specify the relative importance of the criteria. The search circle will then change into an ellipse, if the weights are not equal. For variables that cannot be expressed on a number scale, adapted distance measures can be used.

The economic approach: multi attribute utility theory

Data miners, machine learners and computer scientists often use this idea of distance based nearest neighbor search to match questions to answers or supply to demand. It is not surprising, however, that economists have developed their own concept to model these problems — although phrased differently, the so-called 'multi attribute utility theory' offers a very similar solution.

Assume for instance that we are looking for a second hand car, around 5,000 Euro, preferably red, and blue or green would be second best. For each criterion we can construct a so-called utility function that assigns a utility to all possible

values for the criterion. In our example the utility with respect to price would be 1 for a car of 5,000 Euro, and the utility would decrease for higher or lower priced cars. The utility with respect to color would be 1 for red, 0,75 for blue or green and 0 for other colors for instance. The utility for a given second hand car would be computed by taking a weighted average of these utilities, where the weights would reflect the relative importance of the different criteria price and color. Then instead of choosing the option with the minimal distance to the optimal car (nearest neighbor) we select the car with the highest utility.

Figure 2
Music E-Market (http://www.muziekweb.nl): searching for music by matching with customers with similar preferences.

Matching when there are no product attributes to search on: collaborative filtering

For the sake of the argument, let us stick to the problem of matching demand and supply. There are also markets where matching on a single criteria set of product attributes (color, price, etc.) does not work out. Take for example a home improvement store that offers hundreds types of products, from nails to hammers. A nail is described with different attributes from a hammer, so it requires different matching criteria. Even if the market is constrained to one type of product, only searching on product attributes can make limited sense. On a book or CD market it only makes sense to search on 'title', 'author', or maybe 'genre'; searching on 'number of pages' or 'total playing time' makes much less sense.

A solution to this problem is not to match on product attributes but on purchase histories (also: 'collaborative filtering' or 'recommending'). A famous example is the Amazon.com personal recommendation list. But the example of a music library in Figure 2 may provide a better insight into how recommending works [Putten, 2001]. A customer can list three favorite artists. The recommender

engine starts looking for customers that prefer the same artists and suggests other artists that are common in this customer target group. The counter-intuitive point of this approach is that the engine recommends artists (or albums or songs) without knowing any product attributes such as genre! In other words, the computer can make recommendations with no other information than which customers borrowed which artists.

In the remainder of this chapter we will present two cases to illustrate the data mining approach to the matching problem, with an emphasis on nearest neighbor solutions: catching criminals and hunting for jobs on an electronic marketplace.

Matching case 1: Tracking down suspects

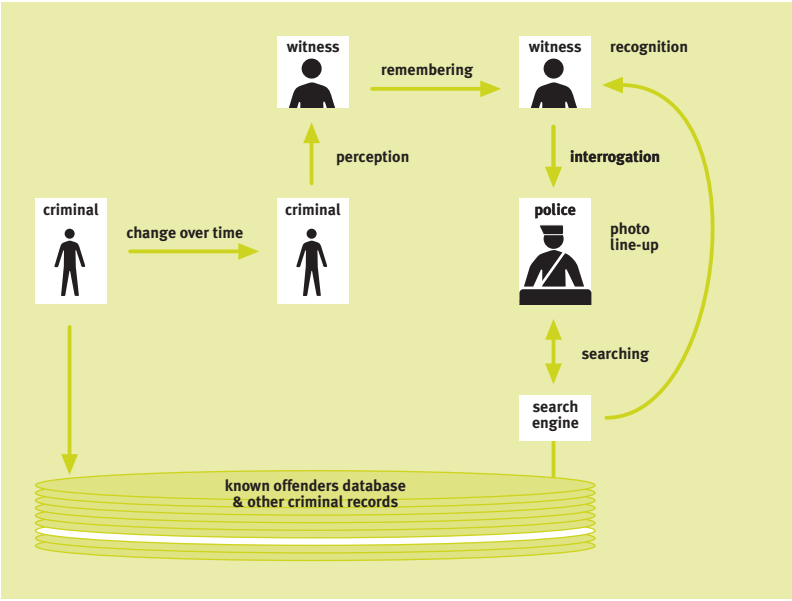
This first example comes from the government sector, more specifically law enforcement. Just like marketeers the police and other intelligence agencies are very much interested in tracking down their 'customers': criminals, terrorists, suspects and victims. It goes without saying that knowledge discovery and matching are core processes for the police and the intelligence community: hopefully investigation results in discovery. On the other hand, forensic data is often complex, polluted, fragmented, error-prone and incomplete. So data mining technologies like nearest neighbor matching can be valuable tools to search and sift through police data.

Business problem

The specific goal in this case is to find known offenders in police data given a description by a witness. The witness is then confronted with a collection of portraits from potential suspects for the purpose of identification. This called a photo line-up or a photo confrontation. Actually, this is a complex multistage process with a high number of bottlenecks and potential sources of error, shown in Figure 3.

The process starts with registration of a suspect or criminal in the police database. Research has shown that there exists substantial difference of opinion between people about descriptions. What is a skinny posture? Where does dark blond hair end and brown hair start? Is this a pale complexion or just normal? Possibly years later a crime is committed and the suspect is seen by the witness. The suspect might have changed over the years, for example he has grown bald, but added a beard or turned from adolescent into a grown-up. Furthermore, perception is an error-prone and subjective process. For example, kids are generally good observers. Characteristics like age and height, however, are likely overestimated. Or the crime has been committed outside in the dark; the witness was stressed and had little time to see the suspect.

Figure 3
The identification process.



Then, with some delay, the witness is interviewed by a police officer; a selection of potential suspects is made and shown to the witness for identification. Again, the memory of the suspect can be distorted, especially after having seen a lot of portraits.

So the police officer that is responsible for making the selection is facing a dilemma. On one side he wants to use all information that the witness has given, which would result in a rich description. On the other side he likes to minimize the risk that the actual suspect is not found, because of a mismatch between the description of the witness and the registration of the known offenders in the database. It is reasonable to assume that the use of standard database queries leads to suboptimal results.

Datamining solution

Nearest neighbor matching provides a way out. The police officer uses all information available and the match engine searches for a preset number of best matching suspects.

An example of such a tool is the DataDetective Associative Recognition System[2] (Figure 4). It was based on the DataDetective Matching & Mining Engine and developed for Dutch police. In the upper left corner the police officer has specified the search query, including weights to express the relative importance or reliability of the individual criteria. In the upper right corner the resulting list of best matching criminals is shown, sorted on match score. In the lower right corner a so-called match analysis graph is drawn, to give an idea of the quality of

2 AHS, Associatief Herkenningssysteem.

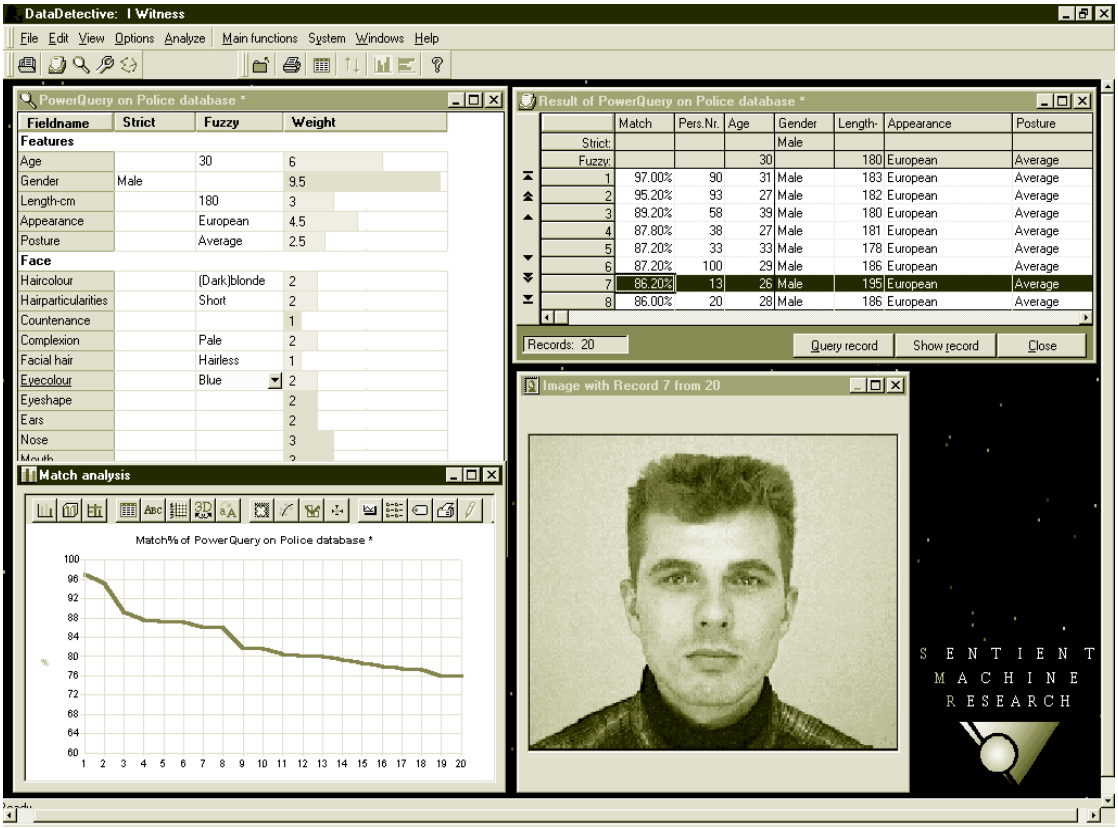


Figure 4
DataDetective Associative Recognition System. Source: Sentient Machine Research.

the selection. It shows the match scores of the best matching criminals found, sorted by match score. So the graph will either stay constant or decrease steadily. From the graph and in the result list can be concluded that no one matched for 100%. In other words, if a standard database query was used, no records would have been found. In this (fictitious) search action the real suspect that was positively identified was suspect number 7 (portrait right corner below), although he did not match for the full 100%.

The matching tool is also used for photo line-ups with a slightly different purpose, the so-called evidence confrontation. In this case the police already suspect someone of having committed the crime. To make the identification task hard for the witness and fairer for the suspect the other portraits in the line-up must resemble the suspect as much as possible. The identification of one white suspect among eleven dark people would not count as evidence in court. By matching on the entire description of the suspect, high quality selections can be made quickly.

The system was evaluated by the police and the following benefits were identified:

Improvement in hit rate. A comprehensive field study was carried out in which crimes were staged and over 150 suspects were interviewed. In this case searching with the matching tool improved the hit rate by 50% compared to performance achieved with conventional query-based selections, from 22% to 34%.
Flexibility. Standard database query constrains the user to use a small number of properties that do not include errors. In an associative search system every bit of information can be specified, without worrying whether the database contains records that match exactly. The more information is used, the better the result.
Ordering by similarity. This is important, because incorrect positive identifications tend to become more frequent, if the witness has already seen more pictures. So it is important to show the best candidates early in the selection.
User friendliness. Previously, selection was the work of specialists and creating a selection of adequate size took a lot of time. With this tool users without special training or experience can search the database; there is no need to develop a strategy to cope with the rigidity of standard database query systems. Actually, the field study showed that users without special photo confrontation training performed better than the specialists! This means that a lot more photo confrontations can be carried out, improving the absolute number of hits even more.

Future prospects

In a police investigation, police officers are constantly trying to match clues to all kinds of known information. Matching tools may be developed that can perform a holistic match on a wide variety of data simultaneously. This might include all kinds of relational and multimedia data such as information on cases, cars, fingerprints, shoe tracks, faces, phone, e-mail, documents, GPS data, DNA and so on. Of course, legal aspects regarding privacy and law enforcement regulations are major issues here.

An example is face recognition (the Pires Face Search & Description System, Figure 5). The tool recognizes a face in a portrait, generates a description and searches for similar ones. The technology behind it is based on a large collection of neural networks and an expert system to generate descriptions and a match engine to match on them. Pires can be used during registration, to generate a description advice and to double-check whether the person registered is already known to the police. The tool can also aid investigation or observation purposes where identification is needed. For instance, it is possible to use a photo composition picture as input. Pictures of unidentified murder victims could also be entered in Pires to find out whether the police know the victim. Finally, Pires could be used in evidence confrontations to select persons similar to the suspect to be identified. This is to make sure that a confrontation procedure is relatively hard for the witness and thus fair.

Figure 5
Pires system. Searching for similar portraits using face recognition and matching. Input can be a photo (left) or a composition portrait (right).

In future forensic information systems we will see more and more 'like-this' functionality in the sense that wherever a user navigates through the various databases in his intranet environment, even if he is using a standard database or keyword search, he will always be able to pull up more cases like this, persons like this, crime scenes like this, etc. You can imagine that a small collection of matching cases and persons is always active in some frame of his screen depending on the context, thus allowing for serendipitous discoveries to happen more often.

Matching case 2: Job matching

Another typical application of matching technology is the job market. The case will describe job matching and mining services offered by Matchcare [Putten, 1999].

Business problem

By describing the job matching case we intend to provide a relevant and frequently occurring example of an E-market. On one side, every organization deals with the selection and retention of personnel. On the other side, people are changing jobs frequently nowadays and are becoming more and more responsible for managing their own careers. Obviously, the friction between supply and demand is an important problem for individuals, organizations and the economy as a whole. A variety of intermediary players try to bridge the gap. For example, government agencies start projects to get the unemployed back to work, headhunters scout for high potential personnel, human resource departments provide job rotation intranets and television, print media and the web are flooded with job adverts.
In reality, the ideal job may not exist (nor does the ideal applicant). And even if it did exist, it might not have been known to the potential employee (or the employer). The Dutch company Matchcare addresses these problems by offering job seekers, employers and intermediaries on the job market solutions

Figure 6
Job market example application. Selection of best matching jobs.

based on web, data mining and matching technology. The main technical challenge is to provide a model that offers enough flexibility to develop solutions for different target groups, but with low costs and the benefits of reusing information and core matching and mining technologies.

Datamining solution

The solution model consists of a small number of core components. All content — jobs and resumes — is centrally stored using a uniform, standardized data model that captures the essence of a job or resume (the 'ontology'). A central match engine provides matching services on subsets of the database. Data mining services are offered to provide knowledge derived from this data. The system can be exploited through various channels and applications e.g. target group specific web sites for the general public, university students , unemployed people who take part in a regional reintegration program, etc.

An example service that has been developed within this model was a job market containing vacancies that appear in Dutch print media.
These vacancies are scanned from magazines and newspapers. Then over two hundred properties may be derived from the vacancies, such as profession offered, industrial sector, type of work, required skills and education, salary range, benefits, contact information, etc. The derivation of these properties is part manual, but also part automated by carrying out text mining on the vacancy and the history of previous vacancies. Some properties are derived using background information such as classifications of industrial sectors and professions that are provided by various statistical and research agencies.
On the web site, consumers may look for jobs for free. Employers may see CV's of job seekers, but only with explicit permission and they have to pay a fee. Job seekers may fill in their resumes with information on education and experience,

build different profiles for each of the target jobs they would like to have, publish their profile, run a match search on the database and apply on-line.

In Figure 6 an example is given. The match engine performs a nearest neighbor search on the resumes and the target job profile. In this case the target job is 'consultant data mining and E-business' and the result includes jobs like consultants for innovation, customer relationship management, business intelligence and E-business. It is possible to apply on-line without giving up anonymity. Before applying, a high-level match analysis can be run to determine the match on the different parts of the profile. This way an applicant gets an idea about stronger and weaker points before the interview. In this example the match was 100% for all that the applicant had to offer: skills, experience and education. However, there was a small mismatch (3%) between the ideal target job and the job offered.

The same job content and core services are used to aid various government and commercial organizations in projects to reintegrate people who are unemployed, for instance because of partial disabilities. In one such a project clients not only receive matching vacancies, but they also receive very detailed information on their strength and weaknesses and suggestions for improvement, given the vacancies they match best for. Data mining techniques such as deviation detection and profiling algorithms can be used for this, in addition to nearest neighbor search.

Vision for the future

So, what can be expected with respect to matching in the future?
First note that in both case examples the actual users of the data mining system were not data miners at all, but rather policemen or site visitors who are probably unaware of the term data mining in the first place. This ties in with an important trend: the democratization of data mining (Figure 7). In the early days, data mining was an apparently magic technology that could only be used by scientists, quantitative or statistical analysts or data miners. However, since 1995, data mining started to become more accessible for data analysis savvy marketeers and business analysts. In future more and more consumers, citizens and business customers will be able to profit from data mining, end users that don't know much (if anything) about data mining implementations or technologies. From a technical data miner's point of view this may not seem important or maybe even not desirable — but just compare the number of data mining analysts with for instance the number of Amazon.com customers and it will be clear what profound effect it will have on the value and viability of data mining in future.
Typically, matching will be on the forefront of this development, because it is a more end-user focused data mining task than clustering or prediction for instance.

If we constrain ourselves to match market — product search applications, there have been some interesting developments over the past years. The first applications were geared towards shop bots for business to consumer sites: think shops and portals for jobs, cars, houses, boats, etc. Then business to business E-match markets (also known as net markets) became en vogue, linking buyers and sellers within a specific vertical: chemicals, plastics, automotive, metals, agro-food, etc. Now businesses become more aware that procurement is actually a strategic activity that shouldn't be shared with competitors, and a new industry is coming up that provides private match marketplaces and auctions for a single buyer: strategic sourcing (E-sourcing) marketplaces. In future there will be room for all three of these application types, each with it's ups and downs. For instance, since the dot.com burst the interest in business to consumer applications has waned, but specialized companies building products such as recommendation systems for video hard disks (Tivo, etc.) are being founded and funded.

There are actually a lot of fruitful cross links to other data mining tasks for matching applications and markets. Prediction for instance can be used for negotiation support: what is a decent salary for me given the jobs I match best with? Is this car over- or underpriced? Clustering and dimension reduction can be used to project the high dimensional space of product attributes into two or three dimensions, thus allowing users to navigate themselves through the space of houses offered for instance, and find the best match.

All of these trends will work towards the same vision: more and more people will be data mining every day — without even noticing it.

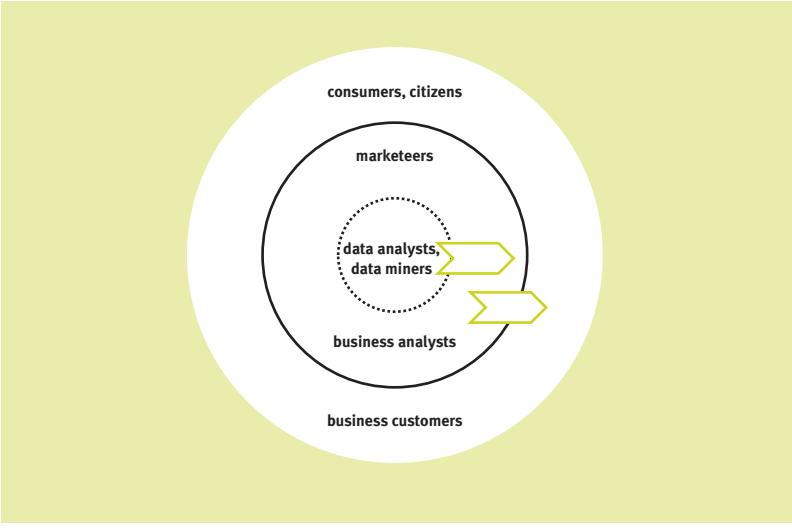


Figure 7
Data mining democratization.

References

- Guttman, R., P. Maes. (1998). Agent Mediated Integrative Negotiation for Retail Electronic Commerce. Proceedings of the Workshop on Agent Mediated Electronic Trading (AMET'98), Minneapolis, Minnesota
- Putten, P. van der. (1999). Datamining in Bedrijf. Informatie en Informatiebeleid **17**:3
- Putten, P. van der, M.J. den Uyl. (2001). Mining E-markets. IT Monitor **3**. Ten Hagen & Stam

Planning of fruit treatment recipes

Floor Verdenius[1]

Background: data mining in agriculture

Data mining techniques penetrate more and more into agricultural domains [GMS Lab, University of Waikato, Verdenius, to appear; Tijskens, 1998; Lokhorst, 1996][2]. Agricultural domains have a number of characteristics that make them an ideal domain for the application of data mining:

- *variable quality* due to cultivar variations, varying growing conditions, varying handling and transport circumstances, etc.;
- *lack of relevant data*: 'hidden' factors that influence quality development;
- *lack of quality knowledge*: unknown factors that influence quality development;
- *evolving quality behavior over time* due to new technologies, cultivar adaptations, evolving market demands and changing quality awareness of customers, and
- *data overload*: modern measurement techniques produce bulk data. Interpretation in terms of quality aspects is laborious and knowledge intensive.

Product quality is a crucial concept for many agricultural products [Sloof, 1996]. Numerous internal processes influence product quality. Internal levels of ripening hormones and their precursors, the exerted post-harvest conditions and handling circumstances determine the ripeness of fruit. Internal levels of hormones, in turn, depend on growing and harvest conditions, such as weather, sun hours, plague pressure etc., and of soil characteristics. Post-harvest treatments (e.g. the application of ripening inhibitors) also may be of influence here. Many of these factors are unknown, either because of lack of fundamental knowledge on product physiology, or because of a lack of adequate measurement technology and information exchange between consecutive actors. Consequently, each batch of produce will manifest its own quality development. For well-engineered products, such as greenhouse-grown products, the variability may be limited. For other produce, for example exotic fruits originating from different countries, the variability may be very large.
All actors in a distribution chain (e.g. Figure 1) make local decisions that aim at both optimizing their profit, and delivering good quality products to the next actor. When non-optimal quality is delivered, the recipient submits a claim for the damage caused. However, non-optimal circumstances at one moment in the

1 Drs F. Verdenius, F.Verdenius@ato.wag-ur.nl, Department of Production & Control Systems, ATO, Wageningen, The Netherlands

2 NNAA workshop as part of ICCTA'96.

distribution chain may only lead to perceivable quality-loss much later. Due to the delayed expression of quality problems, however, quality problems may appear without a traceable direct cause.
In the end, humans perceive product quality. Persons with different roles in a logistic process (e.g. grower, importer, retailer, and consumer) will have a different perception of quality.

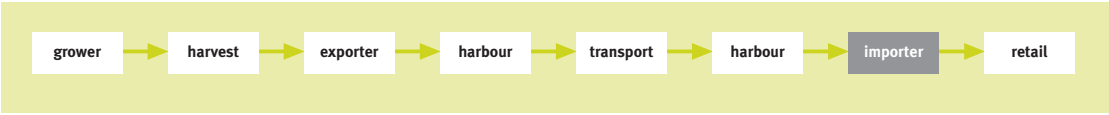


Figure 1
Example of the distribution chain for mango.

For many products, quality experts extract their knowledge of quality development from observations. Based on their experience, they develop methods to assess product quality on the basis of perceivable characteristics. Such assessments are based on visual inspection of the produce. Hidden defects and non-visible aspects (e.g. physiological ripeness) cannot be included in this assessment.
Due to ongoing product innovation and innovations in the technical infrastructure, quality behavior gradually changes over time. Modern cultivars of many products (e.g. tomato, apple) can be stored much longer than some decades ago. Moreover, improved infrastructure allows for better conditions (e.g. fast cooling, smoother temperature behavior during cooling), leading to lower average temperatures, and less quality decay due to temperature variations.
Finally, an important reason for applying data mining in agricultural domains is the management of the ever-expanding quantity of information. Modern development, such as tracking & tracing, imaging techniques and analysis techniques (e.g. micro-arrays, DNA chips, etc.) produce such quantities of data that interpretation has to be automated. Data mining proves helpful in both structuring data (e.g. clustering large quantities of data into relevant groups; e.g. [Broek, 2000]) as in supporting decisions on the basis of measured data (e.g. [Holmes, 1998]). The role of data mining is not limited to specific agro-industrial processes. An example of data mining in cultivar improvement and plant breeding is provided by [Kraakman, 1998].

Motivation for the project

The client in this project is a major importer/distributor of tropical and off-season produce in the Netherlands. Suppliers are found all around the world; the produce is distributed to retailers all over Europe. Constant supply of many products in the assortment is crucial. In practice, at unpredictable occasions quality inspectors are confronted with uncontrollable quality problems. These problems include early ripening of the produce, inhibited ripening, chilling injury, insect damage, (fungal, viral, bacterial) infections and other quality problems.

For most products, storage and transport conditions are standardized. These standards are based on the average product quality. Intra-batch quality variances, variances between 'identical' products (products of same cultivar, size, origin) in the same shipment, are partially responsible for quality problems. Inter-batch quality variances, variances between different shipments, are another main problem. It is the experience of product experts in the client organization that 'identical' batches behave differently under 'identical' conditions. This is caused by the fact that 'identical' in the eyes of the importer is not 'identical' in terms of physiological processes.

Client requirements

The goal of the reported project is to assess the potential value of an approach where, on the basis of systematic (and as complete as possible) information processing, the importer can be supported in:

- assessing the maximum period under optimal conditions that the product will have a minimum required quality;
- planning optimal conditions for a product to have specific quality at a specific instance in time.

Additionally, the approach should be applicable to any product. Mango serves as the pilot product.

Project approach

The aim of the project is to collect quality-relevant product information from the entire life cycle of a product batch, and to interpret this information for planning of future product handling. Product information accompanies a batch downstream. Every link in the distribution chain, every process or treatment, adds its quality-related information to the information stream. When an actor wants to assess the quality of a batch, a local quality snapshot is completed with the available information to derive the quality of a batch.

To determine the quality downstream, it is crucial to collect information as early as possible. Upstream actors, however, will not provide that information (on their business process) for free. Therefore, upstream information that enables upstream actors to improve their business processes in future is as crucial for project success as is the downstream flow.

In the project, three goals have been pursued in parallel:

- Definition of quality relevant information throughout the distribution chain.
- Organization of efficient and sufficient information flows down- and upstream.
- Realization of a prototype planning system that supports product experts in designing treatment recipes for the above mentioned planning tasks (determining maximum duration under optimal conditions and determining a recipe for a specified quality at a specified time).

Project results

This text does not elaborate upon the first two points. It suffices to notice that concerning the functionality of the result, the definition of quality-relevant information has delivered a detailed model of what mango ripening implies. In its most fundamental physiological form, this results in interesting but technically immeasurable assessment criteria for ripeness. In a simplified form, it delivers a model that can be used to assess ripeness of a batch, given initial state and relevant conditions (and so it also defines what those relevant conditions are). Useful for determining the quality given a recipe, the model is unsuited for defining a recipe, given an initial state and a quality requirement. In other words, it can be used to 'assess the results of a plan', not to 'design a plan'.
The quality-related product information that accompanies the product flow has been organized for a grower-exporter-transporter-importer combination. At harvest time, the grower records information on product quality, growing conditions and harvest and post-harvest handling. From the beginning of the distribution chain onward, storage conditions are logged, and periodic quality assessments are performed. On arrival at the importers facilities, a batch is accompanied by a substantial quantity of information. Apart from the process information, this includes administrative information concerning origin, product and handling, and batch characteristics such as packaging and shipment type. All this information is entered in the Treatment Advice System (TAS) that facilitates storage-recipe planning for the product expert.
Figure 2 shows the schematic set-up of the TAS. Product flow enters the assess step. A quality measurement is performed, and sent together with a batch recipe specification (deliver quality q at time instant t) and all available historic and administrative information on the batch to the next activity. In the specify activity, available information on the new batch including the recipe specification is matched against information from historical cases as stored in a database. The recipe for the best matching historical case is then used as ballpark recipe for the new batch. Based on a detailed comparison between the new batch and the best matching historical case, the ballpark solution is now adapted to optimally suit the new batch. This results in a global specification of the recipe. The global specification is a compact representation of recipe characteristics, such as the quantity of heat to transfer to the product, the quantity of ethylene gas, the relative humidity, etc. that have to be applied for successful delivery of results. In the plan process, the global recipe requirement is converted into a detailed recipe that can be used as process set points. In the validate step, the application of the detailed recipe on the batch is simulated with a product quality development model. This model gives a tentative validation of the recipe. If undesired effects emerge in the diagnose step, this triggers the adapt step to compensate the recipe for undesired effects.

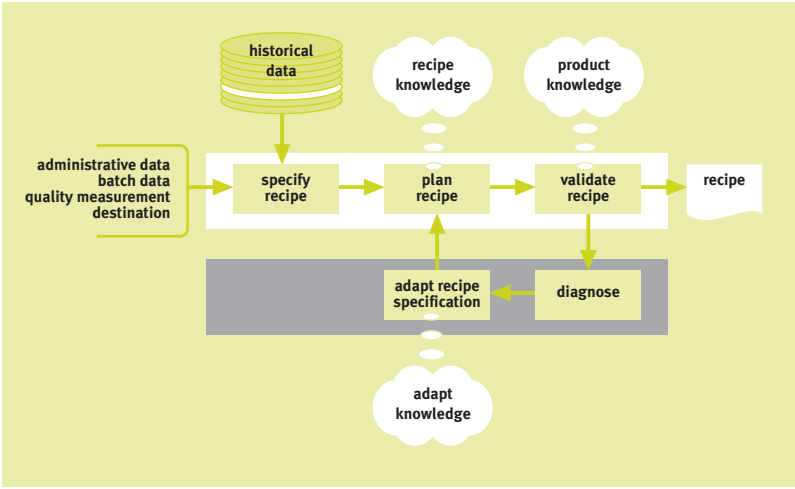


Figure 2
Schematic set-up of the Treatment Advice System (TAS).

Future developments

Introduction of the TAS prototype in practice is currently taking place. Initial experience shows that building up the case base is time-consuming. For the moment, major benefits are an improved communication with partners in the supply chain, and an improved perception of quality from regular use of the TAS. This improved quality awareness leads to better monitoring of quality development, and early warning for quality problems.

Practical experiences also indicate several options for improvement. Matching and search criteria should be made more flexible, while their results should fit better within existing heuristics. Moreover, the response time for planning is too large. All these criticisms, however, hardly affect the case-based core of the system.

The desire of the client to improve quality management requires adaptation of the activities in handling fruit. Currently, the storage approach is reactive: commodities are stored under standard conditions, and sold in small parcels. Based on regular quality inspections (of cosmetic features), storage conditions can be altered and commercial strategies for that batch (price, client destination) can be adapted. However, when external features express quality loss, the remaining shelf life time is short.

Implementation of TAS in practice facilitates pro-active business processes. The whole process applied here is often referred to as case-based reasoning. The system currently runs for one product, and a limited number of supply chains. In principle, expansion to other chains and products is possible. A major problem, however, is organizing a complex data collection process over a large number of actors with substantially different levels of expertise, technical development and product expertise. For educating product experts worldwide a simple and accessible education schema has to be developed. This needs to ensure a correct and uniform information content. Moreover, a global informa-

tion-processing infrastructure is required; Internet implementation is an option. The role of learning techniques in the TAS follows from the non-availability of reliable product models that can optimize for storage and treatment conditions. In the absence of substantial product knowledge and methods to objectively assess physiological ripeness, exemplary data is the only source for generalizing recipes from batch specifications. In the planning schema of Figure 2, the match, adapt and plan activities represent one possible realization of planning competence. A preceding version for planning treatment recipes [Verdenius, 1996] used a neural network to assess global recipe requirements, and had less facilities for feedback on recipe quality. The basic approach was similar. The TAS approach, however, offers a better integration of existing knowledge and experience with the inductive power of data mining techniques. Available knowledge on product quality development is used for validating the recipes. For some agricultural products, quality development models may be developed that facilitate proper planning, replacing the CBR planning module with some model-based planning module. For many products, however, availability of such models is not to be expected.

An interesting challenge would be the combination of model-based and data mining approaches, resulting in the induction of mechanistic models for ripening. Given the proliferation of tracking and tracing needs for agricultural produce, and the expanding role of Internet as communication medium, one development that can be foreseen is that product data become available in large quantities. This would enable a systematic approach to be implemented against relatively low costs. One future development therefore is the transformation of the TAS to an Internet based system. This will not so much influence the data mining content, as well as fulfilling a major condition for large scale proliferation of this kind of systems: availability of data.

References

- Broek, W.H.A.M. van der, J.C. Noordam, A. Pauli. (2000). Multivariate Imaging for Automated Process Control in the Agro Industry. In: Proceedings of Agricontrol 2000. International Conference on Modeling and Control in Agriculture, Horticulture and Post-Harvested Processing. To Appear
- GMS Lab. http://www.gis.uiuc.edu/cfardatamining/default.htm
- Holmes, G., S.J. Cunningham, B.T. Dela Rue, A.F. Bollen. (1998). Model-IT Conference. Acta Horticultura 476. Edited by L.M.M. Tijskens and M.L.A.T.M. Hertog, The Netherlands. pp289-296
- Kraakman, A. (1998). Application of Machine Learning in Plant Breeding: The Process of Putting a Learning Technique into Practice. In: F. Verdenius, W.H.A.M. van den Broek. (eds.). Proceedings of Benelearn 98. pp148-156
- Lokhorst, C., A.J. Udink ten Cate, A.A. Dijkhuizen. (1996). Proceedings of ICCTA'96. Agro Informatica Series nr. **10**. Wageningen

– Sloof, M., L.M.M. Tijskens, E.C. Wilkinson. (1996). Concepts for Modelling Quality of Perishable Products; Trends in Food Science and Technology **7**:165-171
– Tijskens, L.M.M., B. Nicolaï, M.L.A.M. Hertog. (1998). Proceedings of the First International Symposium MODEL-IT. Wageningen, HIS. Acta Horticultura 476
– University of Waikato. http://www.cs.waikato.ac.nz/ml/
– Verdenius, F. (1996). Managing Product Inherent Variance during Treatment. Computers and Electronics in Agriculture **15**:245-265
– Verdenius, F., L. Hunter. (2001). Pros and Cons of Inductive Modelling. In: L.M.M. Tijskens, M.L.A.M. Hertog. Food Process Modelling. Ellis Horwood Publishers, Cambridge

Towards a self-adapting insurance company

Charlotte Bouvy[3]

Introduction

The new possibilities of data mining, or Knowledge Discovery in Databases (KDD)[4], have not gone by unnoticed in the insurance business. In this chapter I will address some forecasts for the use of data mining in the insurance business. Making predictions always bears the risk of revealing more about the present and the recent history than about the future. That counts for this article as well. I will start with a short overview of the present situation and how we got there.

Recently, Dutch car insurers presented unexpectedly bad results: a 10% higher claim value than last year, with only 5% an increase of premium income. The business of an insurance company comes down to risk assessment: to make good estimations of the future claim behavior of clients, their number of claims and the claim value. An insurance company that knows its risks can make better underwriting decisions: this company will underwrite the acceptable (low enough) risks and reject the higher risks, or add a premium factor for the higher risks, while calculating a more competitive premium by offering reductions for the low risks. This company will make profit.
So, the challenge is to define a model that assesses the 'risk score' or the accept/reject decision, when a client applies for insurance. Other examples of models are future claim behavior models or future loss estimation models, models that make estimations about an entire policy portfolio or that define distinct risk groups within that portfolio. A more advanced risk model will give better results.
After this, the next challenging step will be to adapt this model to the ever changing world, to learn from the customers.

3 Drs M.C. Bouvy, Charlotte.Bouvy@bolesian.nl, Knowledge Analyst at Bolesian (member of the Cap Gemini Group), Utrecht, The Netherlands. http://www.bolesian.nl

4 The use of Knowledge Discovery in Databases for data mining emphasizes the 'discovery' aspect and the focus of discovery on 'knowledge' (whereas many statistical methods are more of a 'hypothesis confirmation').

Risk assessment requires knowledge

Risk assessment is a knowledge intensive operation and requires explicit representation of knowledge: for example a rule set or decision tree expressing the business rules or acceptation norms [Schreiber, 1999]. The explicit representation of knowledge — rather than having it intertwined into the procedural code of a computer system — makes it possible to maintain, update or adapt the knowledge. Moreover, knowledge in an explicit representation makes it possible to motivate a decision.

Rule-based risk assessment provides an objective and consistent evaluation of acceptance criteria. In fact, by using business rules in the front office, the client is only asked for information that is relevant in his or her situation (in stead of asked to fill out a tediously long application form). This front office can be anywhere: a department in the insurance company, a call center, an insurance agent, all using the same business rules.

Traditionally, the knowledge needed to build a knowledge intensive system[5] is acquired from human experts — for the insurance domain typically a senior underwriter or an actuary. Machine learning techniques made it possible to discover knowledge in data, a source that can complement human expertise. Where human experts are excellent in recognizing patterns from fuzzy data and reasoning with incomplete data, they do have some fallacies, when it comes to analyzing data.

Strange relationships

Human assessment of risk factors tends to move along linear or simple relationships such as: 'the bigger the company, the lower the risk', or 'male drivers have a lower risk of car damage than female drivers'. When variable factors not only influence risk, but also each other, the overall effect is harder to foresee (although an insurance expert will know how to explain the combined effect of age and sex on car damage risk).

The multiple and variable factors involved in such a complex issue as 'risk' have an impact on each other in ways that are often too complex for a human insurance expert to process and analyze. A high correlation between variables does not always explain in terms of cause and effect. Possibly a third hidden factor has to be discovered to explain the findings. For example, a human observer might find that there is a high correlation between a high 'number of claims' and a 'payment method' in an insurance clients database. A risk assessment model based on the payment method the client chooses on the application form is clearly not desirable! Data mining this database can reveal the hidden variable that could clarify in terms of causal relations: the relation with 'type of coverage' [Holsheimer, 1997-1999]. For example, one might discover that Hailstorm insurances are usually paid for differently than third-party insurances.

5 Decision support system, expert system, business intelligence are all knowledge intensive systems.

Large quantities of data

Decisions in the insurance industry are often based on data which is not only far too complex, but also far too voluminous for the human mind to process and analyze. Consequently, decisions are likely to be based on the fallible human observation of a small subset of a large body of relevant information. Making the best decisions is even more challenging, because the insurance business is transaction intensive and still lacks the extensive electronic tools needed for managing and tracking all relevant information.
Neither complexity nor massive quantities of data pose problems for the modern data mining techniques.

Quantity versus richness

Pilot studies in recent history have shown that integration of risk assessment into operational systems is still a great challenge. The same goes for applying the knowledge that is discovered in data. Why is it that these data mining techniques that have been available for over a decade aren't embedded in every insurance process yet?
Traditional (or today's) insurance companies have a long automation history, characterized by many stand-alone back office systems, mainframes, archives, cabinets with dossiers. Data files are fragmented and for a large part only available on paper. Data is collected, organized and stored for insurance process purposes: administration of policies, payment of claims, calculating premiums, etc., clearly not for learning-about-risks purposes. Many of the systems were designed in an age where storage required physical space. So, although there are huge quantities of digital data available, it holds very few risk factors. This is not to say that insurers do not know their clients. They know them very well. These risk factors are abundant in the application forms that clients filled out and in the case of business clients in the paper dossiers of inspections.
Only recently, front offices started to use risk assessment systems in the application phase of the insurance process. Since then that data on 'soft' characteristics of clients has been used and even stored. Add to this the possibilities of enriching with external data (regional- or zip code-based characteristics or data collected by branch organizations for business insurers) and it becomes clear that a lot more can be done.
In a few years paperwork will be history with electronic application forms and web-based risk assessment systems emerging on Internet.

Data mining requires domain knowledge

A second reason why data mining still has to take that flight in the insurance business is that a technique or tool sometimes falls into the wrong hands: those of technicians. In other words: one has to know the business in order to benefit from data mining. The people knowing the issues, challenges, problems and

threats of the insurance business are not those with a special interest in data mining techniques.
This works in both directions: a machine learning technique will give better results, if the data it has to learn from is 'biased' with domain knowledge or even common knowledge. In a dataset with over 90% of 'reasonably low' A-risks and only 0,5% of 'extremely high' D-risks, one needs to tell the technique that to mistake a D for an A is more harmful for the company than the other way round. Only an insurer can decide which factors may appear in a risk assessment model (remember the example of the model that bases its predictions on the clients' payment method).
When learning from examples, it should also be kept in mind what and who is being learned from: in this case often only data on policies that are already in the portfolio, and these are the risks that were once assessed to be eligible! Nevertheless, there is a lot to be learned: adapting to changes within this group, predicting new risks in this group of policy holders, recalibrating the bottom line, estimating future losses, tuning the risk factors in the premium calculation, detecting the extremes. Insurances have a collective character (or a solidarity aspect if you wish). Therefore, groups that are just below or above medium risk are not interesting. Newly emerging risk factors, changing influences of risk factors, and extreme deviations are specific events to be discovered.

Not every model resulting from a data mining analysis makes sense. The goal of a model must be clear beforehand: should it offer a premium reduction to lower risks or reject extremely high risks? The purpose of the model directs every step in the data mining process, and only with a clear goal in mind an insurance expert can interpret and test the findings of a data mining analysis. The actuarial and legal consequences of a model must also be evaluated. Here lies the real challenge: to integrate the knowledge of the expert with the knowledge discovered in data.

Tomorrow: the adapting insurance company

Now that insurers see that they should also store rejected applications because they can learn from their data, we can look forward to the insurance company in, say, 10 years from now. It will have its business knowledge represented in several models for different purposes, explicitly in the form of rule sets or decision trees. The models will be kept up to date on the basis of the actual data, but will always result in risk models that stay within the actuarial and legal boundaries. These models will be used not only to assess the risk of new clients when they apply, but also periodically to analyze the lifetime profitability and risk in the existing book of business and to use this analysis to forecast future risk for the policyholder. This company is a pro-active financial planner for its clients. It

knows, when it is time to offer an extension on the car insurance coverage for the son that just turned eighteen.

References

- Holsheimer, M. (1997-1999). Data Distilleries. Several Articles Published in Informatie
- Knowledge Stream Partners. gps@ksp.com
- Piatetsky-Shapiro, G. Knowledge Discovery in Databases: 10 Years after
- Schreiber, A.Th., J.M. Akkermans, A.A. Anjewierden, R. de Hoog, N.R. Shadbolt, W. van de Velde, B.J. Wielinga. (1999). Knowledge Engineering and Management: The CommonKADS Methodology. The MIT Press, Boston, MA

3

3.3 Conclusions and Expectations

Jeroen Meij

This chapter will provide some conclusions and expectations for the future based on the cases from Part 3. Furthermore, some of these expectations are illustrated with two fictitious cases of future companies. The last section is a short summary in key phrases.

3.3.1 General

For business users, government users and consumers alike, data mining is expected to move from the domain of the specialist (data miner, statistician, scientist) into the domain of the user. First reaching the business analysts and marketeers, then reaching business and private customers. The democratization of data mining as a process is just starting and will continue. As with many other successful developments, data mining tools will be embedded in a wide range of software products and services, often without the end user realizing it. With progressing virtualization of business (see below), the need for data mining tools grows, and this will ultimately lead to real time contextual adaptation.

Customer relations management

Data mining will become an integrated part of the marketing process within many companies. Creating customer profiles is likely to become a more complex and specialized activity. Companies will store information about failed sales attempts as well as successful ones. Business knowledge will be represented in several models for different purposes, explicitly in the form of rule sets or decision trees. The models will be kept up to date on the basis of the actual data, but will be watched closely to keep them within the actuarial and legal boundaries.

These models will be used not only to assess the potential of new customers when they apply, but also to periodically analyze the profitability and risk of the existing products and services.

For the consumer, the companies will use analysis tools to forecast future risks or desires. Companies will move toward pro-active need planning for their customers, knowing when to offer an extension on the car insurance coverage or when to offer which new product.

It seems reasonable to assume that the trend towards more personalized and ultimately one-to-one marketing will continue. As public awareness on privacy issues increases, companies that perform marketing on a non-selective, wasteful manner risk being neglected by customers. Even so, one-to-one relationships must be meaningful from the perspective of the customers as well.

Ultimately, good data mining practice — involving ethical as well as commercial principles[1] — could lead to benefits for both the selling companies and the customers: less undesired sales contacts, and a higher percentage of welcomed sales contacts.

Customer emancipation is imminent: when the appropriate software tools (or services) are available, intelligent agents will scour electronic markets, representing individuals or groups of customers to search for necessary, interesting and useful products. These customers will only release profile information, when they feel it is to their benefit.

1 See Chapter 4.1, Web mining in a business context.

Business action related conclusions

Segmenting

Segmentation is a well-known method in customer relations management, and will continue to play an important role in this area. Other usages will be common as well, such as insurance risk analysis, policy research, market basket analysis, crime analysis and medical discovery.

A shift is expected from univariate profiles (relating one variable to another) to multivariate profiles (many to one) or non-propositional profiles (many to one or many). The multivariate profiles could be generated by association rules and decision trees, the non-propositional profiles could be constructed by inductive logic programming algorithms. For segmentation purposes, visualization technology such as three-dimensional segmentation and visualization is expected to play a more important role.

Data fusion methods could allow us to enrich entire customer databases with survey information that is only available for a sample, in other words, carrying out a virtual survey with each customer [Putten, 2000]. If this technology becomes mature, a whole new arena for segmentation will evolve.

Classification

Standard classification depends on a supervised learning process, which indicates representative examples for the different predefined classes. Often this involves labor intensive manual selection. Developing automated adaptation systems will be a logical next step. This could involve incorporating a segmentation (see Section 3.2.1) step into the process that only requires supervision to be linked to the desired classification. This adaptation can be done on request or periodically. Systems such as these are expected to be used in many areas, quality inspection being an area of particular interest.

Detecting

As data collection increases, so does the importance of automated detection. In many cases it is sufficient to know when human interference is required, and automated detection is a cheap alternative to human observation and analysis. Many applications can be thought of, for example; traffic supervision, monitoring of public spaces, fraud detection, intrusion detection (in the real world and in computer networks). Needless to say, advances in this field are closely related to developments in pattern analysis and recognition. An interesting question for the near future is whether people prefer to be monitored by humans or by computer systems.

Modeling

Creating models will help us to understand processes, patterns and behavior. With advanced data collection techniques, more and more reliable data become available, allowing us to create better models.

Prediction

A second step after creating a model is to predict the effect of input variables. Better models allow for better prediction. Eventually, this will lead to user-friendly decision support systems.

Matching

Matching is one of the first areas where the end users themselves will be using data mining techniques. E-markets are a vital element in this development. We will see data mining — combined with agent technology — playing an important role in many matching and linking situations. This applies to business to business, but also to business to consumer and consumer-consumer relations.

Adapting

Self adapting systems and services are among the more advanced developments that will be seen in the near future. We can expect self adapting systems for many of the tasks described in this part of the book. These systems will adapt themselves to new conditions, products, quality demands etc. of a process. At the far end of the scale is immediate adaptation to the actions of a customer on a web site, proposing tailored offers or information based on continuously updated models.

3.3.2 Future cases: data mining in virtual organizations

Lex Kwee[2], *Walter Kosters*[3]

Introduction

In this article we describe two cases of future companies. These companies are virtual in two or even three aspects. First, in having a virtual business model, in which all non-core activities have been outsourced to specialist providers. Secondly, in being active in electronic business wherever possible, reducing physical operations to a minimum by making maximum use of information technology. Finally, the second company does not sell physical products. Its business is based on bits and bytes of intellectual property that can be distributed electronically. The only physical product aspect is packaging to make the core product more accessible.

2 Drs A.Y.L. Kwee, lk@zi.nl, New Business Associates, Abcoude, The Netherlands

3 Dr W.A. Kosters, Kosters@wi.leidenuniv.nl, University Leiden, LIACS, Leiden, The Netherlands

The virtual organization

The concept of the virtual organization was launched in the early nineties. Davidow and Malone described how companies focus on providing the ultimate customer service by applying organizational concepts such as mass customization in a virtual business model [Davidow, 1992]. Mowshowitz presented a theoretical framework for business design and operational decision making in virtual organizations [Mowshowitz, 1994; Mowshowitz, 1997]. Both approaches depend on the availability of large quantities of operational data and explicit business rules that transform operational data into planning recommendations and fulfillment execution. It will be evident that a virtual organization depends heavily on data mining.

Supply chain outsourcing

A virtual organization is set up around selected groups of customers. Every activity that does not directly contribute to the customer's experience and satisfaction may be outsourced. Examples can be found at non-manufacturing companies such as Nike and Ikea, which focus on product design, marketing and distribution, leaving actual production outsourced to low cost providers, and, in the case of Ikea, rely on the customer for final product assembly. It is predicted that traditional manufacturers will increasingly adopt this approach, as can already be seen in the automotive and IT hardware industries, in which the traditional manufacturers have outsourced the majority of manufacturing work , turning their traditional factories into final assembly shops and warehouses. Some companies, such as Hewlett Packard, have even outsourced final assembly and warehousing, making it part of the delivery process, which can be outsourced to global logistics providers and supply chain fulfillment companies. A similar approach in retail was pioneered by Wal-Mart, that sends cash register sales data directly to its suppliers and allows them into its stores to restock the shelves. This type of strategic outsourcing is dependent on highly integrated design, product management and supply chain planning systems, in which data mining applications constantly monitor behavior according to preset schedules and exception criteria.

Collaborative electronic commerce

The latest virtualization trend is the rise of collaborative electronic commerce, in which trade partners cooperate to serve markets, manage suppliers and logistics and implement electronic match making mechanisms, using a shared external ICT infrastructure and a common marketplace service. All these activities are highly data dependent, requiring a high degree of system and process connectivity. The increases in possible trade relations and electronic transaction volumes generate significant operational complexity. Companies need to connect, exchange and translate information electronically, providing instant feed-

back and situation specific pricing, terms and conditions. This forces companies to align processes and adopt emerging standards to achieve the benefits of fully automated collaborative electronic commerce. End customers can interact with the business systems of all players in a supply chain, leaving valuable data traces, which can be used for future marketing and customer service purposes. As the level of supply chain cooperation increases and the volume of electronic commerce rises, the volume of search, transaction and commerce related data will explode. Market tactics and business operations are becoming increasingly dependent on data mining.

Case 1: automated supply chain

A corporate Purchaser usually buys from a pre-selected set of suppliers. When a good offer is found at an electronic marketplace, an order is initiated. The information is automatically sent to an electronic supplier clearinghouse. The clearinghouse transaction hub:

- applies the Purchaser's business rules to match the supplier profile to the Purchaser's criteria;
- translates the order to a processing standard that is used on the shared supply chain management system;
- performs a credit check and
- sends the information to the selected supplier.

The supplier receives the order and sends a confirmation, via the hub, where it is translated and forwarded to the Purchaser's planning system. Given the value of the order, a business rule instructs the transaction hub to send a notification to the Purchaser's mobile phone for confirmation.
Fulfillment is monitored automatically and when the predefined situation is reached, another business rule triggers the transaction hub to send an email to inform the Purchaser that the order is ready to be shipped and that all export documents have been cleared. If no exceptions occur, a payment transaction is triggered automatically.
During these transactions, a lot of data is produced. Mining these data will expedite future transactions. The Purchaser will combine process and fulfillment data to manage vendor rating and volume purchase contracts. The Supplier will combine process and payment data to manage a buyer rating and update pricing policies. Unexpected patterns will be used to activate exception handling and notification rules. Purchaser and Supplier may decide to outsource data mining to the marketplace, which has already outsourced this to the transaction hub. Demand analysis and forecasting will be used to determine optimum levels of stock, spare parts and production capacity.

Case 2: Multimedia distribution

Background

A global multimedia conglomerate deals with all aspects of entertainment. The firm is large, and its activities take place throughout many different countries. In addition to internally produced entertainment goods, the firm sells external products, and uses external providers for specific technologies (e.g. video games). The firm's core business in this situation is branding. Distribution of media is outsourced, but all logistic and service data are collected for internal analysis.

In this approach, the firm does not know its final customers. Therefore, it makes heavy use of coupons and web forms, encouraging consumers to exchange personal data for discounts.

Also, the firm has a live entertainment division that provides customers with the opportunity to interact with the firm and its products in an attractive environment.

Future

Logistics

For an entertainment company, logistics are crucial. If products are ordered using the Internet, both the customer and the firm need to keep track of the status of the order and its payment. As many intermediaries (banks, logistics providers, warehouses) are usually involved, order tracking used to be a major problem. Agents can take care of this by searching the appropriate databases (which can be of totally different structure), extracting the knowledge about the product at hand, and presenting this to both customer and company. Data mining will predict logistic bottlenecks (by examining records from the past) and will come up with alternatives (by examining the existing resources). Finally data mining provides tools to deal with fraud. Data from transactions from the past can be used to detect fraudulent patterns.

Improved techniques allow for 'just in time' delivery on all levels. Since many products are digital, network behavior and server load prediction deserve attention. New techniques combine recent and expected behavior into on-line monitoring of the whole process. This may facilitate dynamic pricing, again based on customer observation. It is expected that in many industries pricing will follow the dynamic revenue enhancing approaches that have been developed for airway companies and have led to the myriad of price variations for simple airplane tickets. All available parameters are used to find a match between the buyer's willingness to pay a price and the seller's risk of having to fly with empty seats.

These approaches rely on heavy computing power and can support a limited number of transactions per second.

Marketing

The main area where modern data mining techniques can be applied is marketing. Here the questions are quite diverse. One may ask for customer profiles, but possible connections between purchases and predictions of sales are also interesting. A database of past transactions is extremely useful for customer profiling, in combination with the web-log database this may yield powerful tools. It may also be possible to automatically combine different databases (data fusion).

It may be imagined that adding computing power will lead to a situation in which the user's information environment continuously adapts, maintaining a personalized commerce space in which multiple offers are available, and adapted in real time. This will not only include personalized pricing, but also personialized delivery, packaging and even personialized products, in a future extension of Dell's seemingly 'made to order' PC assembly model or Unilever's Rituals 'made to order' cosmetics approach. Soon personalized shops will be available on the Internet for all types of configurable and customizable products. New graphical techniques will be used to spice up the presentation.

Amidst these, mainly technology driven, 'enhancements', the high level goal is a consistent and meaningful view of customers. For this purpose all existing databases need to be examined and their contents should be combined. Existing models will be further developed and improved using extensive calculations. On the one hand black box methods get better and better, on the other hand the management stresses the importance of understanding behavior. Location specific personalization, using inputs from mobile communications and computing devices, will result in real-time situation specific personalization. The entertainment company will not send its offer for a comprehensive vacation package (entertainment included) to a user's mobile phone during a car drive in an unknown inner city. The offer will be held until the prospective customer has reached his hotel, and switches on the television, where he can watch a highly personalized version of what his next vacation could look like. The trigger for this specific offer came after mining location data from his mobile phone. If he decides not to switch on the television, the phone will point him to a bookstore, museum or restaurant, all in accordance with the customer profile, preferences, situation and location.

When consumers prefer to be entertained at home, suppliers will be ready to serve their home entertainment centers. Entertainment on demand such as games and virtual trips in space can be provided, using equipment such as body suits and 3D visualization. These techniques offer a lot of opportunities for the entertainment business, but also raise a lot of questions concerning logistics, customer profiles, privacy, etc. Participants in such a virtual reality game or trip want to be amused in a way which fits — among other things — their emotional state and personal feelings. In order to determine the emotional state one has

to investigate the behavior of the customer. The virtual entertainment center gives a lot of information on this behavior; the physical reactions and the reflexes of a person to a certain situation, how the person proceeds, alterations to the scene, and all sorts of other information. All this information is, of course, encoded in raw data. Data mining can be used to analyze, decode this enormous amount of data and to determine the part of the current customer profile relevant for the virtual entertainment of the customer. The books or CDs a person buys can (and must) also influence the profile, since it gives information about the kind of entertainment a person likes. The ending of a person's favorite stories gives relevant information on the endings for the virtual entertainment games. So one also has to mine this data in order to get the very specific customer profile needed to offer interactive entertainment that fits the desires of the customer. This customer profile together with the perceived temporal emotional state of the participant in the virtual entertainment session can be used to determine the continuation of the virtual interaction in such a way that the participant is very satisfied at the end of the interaction. In this way personalized entertainment is within reach. On the other hand the behavior in the virtual entertainment center can be used to extend the customer profile for the more classical channels of entertainment, and makes it possible to offer certain (electronic) books or CD's that correspond to the behavior in the virtual world.

Accounting

In business areas where available data are more complete and precise, such as accounting and financial reporting, data mining may be used in semi-autonomous mode, for instance, in decision systems. In areas where data is scarce, incompatible and or unstructured, human supervision and analysis will be needed.

One can expect that the methods mentioned above will be extended and improved. In a sense the techniques might still be called classical, since the main new ingredient is computing power.

Related issues

On the basis of these previous examples of future companies, we may predict several issues that deserve our attention.

Privacy

The enormous quantity of customer data is unfortunately prone to politically incorrect approaches. It is also clear that border crossing activities, especially when dealing with information in general and entertainment in particular, may lead to unwanted behavior. For this topic we refer to Part 4.

Unclear pricing mechanisms and information overload

We expect to see many autonomous competing firms, who are eager to use novel technologies in a rapid way. Technology will result in new combinations of product, packaging, personalization and delivery. Pricing will have to reflect these new possibilities. Some authors argue that the decisive factor in buying decisions will not be product-based. Rather, experience (immediate, user controlled delivery) will be the determining factor that sets new price levels. In addition to the pricing models of airlines, companies will have to adopt menu and serving suggestions from restaurants and other leisure companies.

As a countermeasure against any possible information overload driven by data mining, consumers can use data mining themselves, to compare products and prices, or they can use tools like I AM [4] to put their personal preferences in control over external stimuli and inputs. And finally, consumers will adopt multiple (electronic) personalities to allow them to continue to browse, buy and consume anonymously where they prefer to do so.

3.3.3 Closing remarks

To make a very condensed summary of the conclusions, we expect the following trends to materialize:

- democratization of data mining;
- integration of data mining in many business processes;
- automation of adaptation cycles;
- agents aiding the emancipation of consumers.

Aspects like ethics, privacy and solidarity are important issues to be discussed in this context. Part 4 of this book is dedicated to these matters.

References

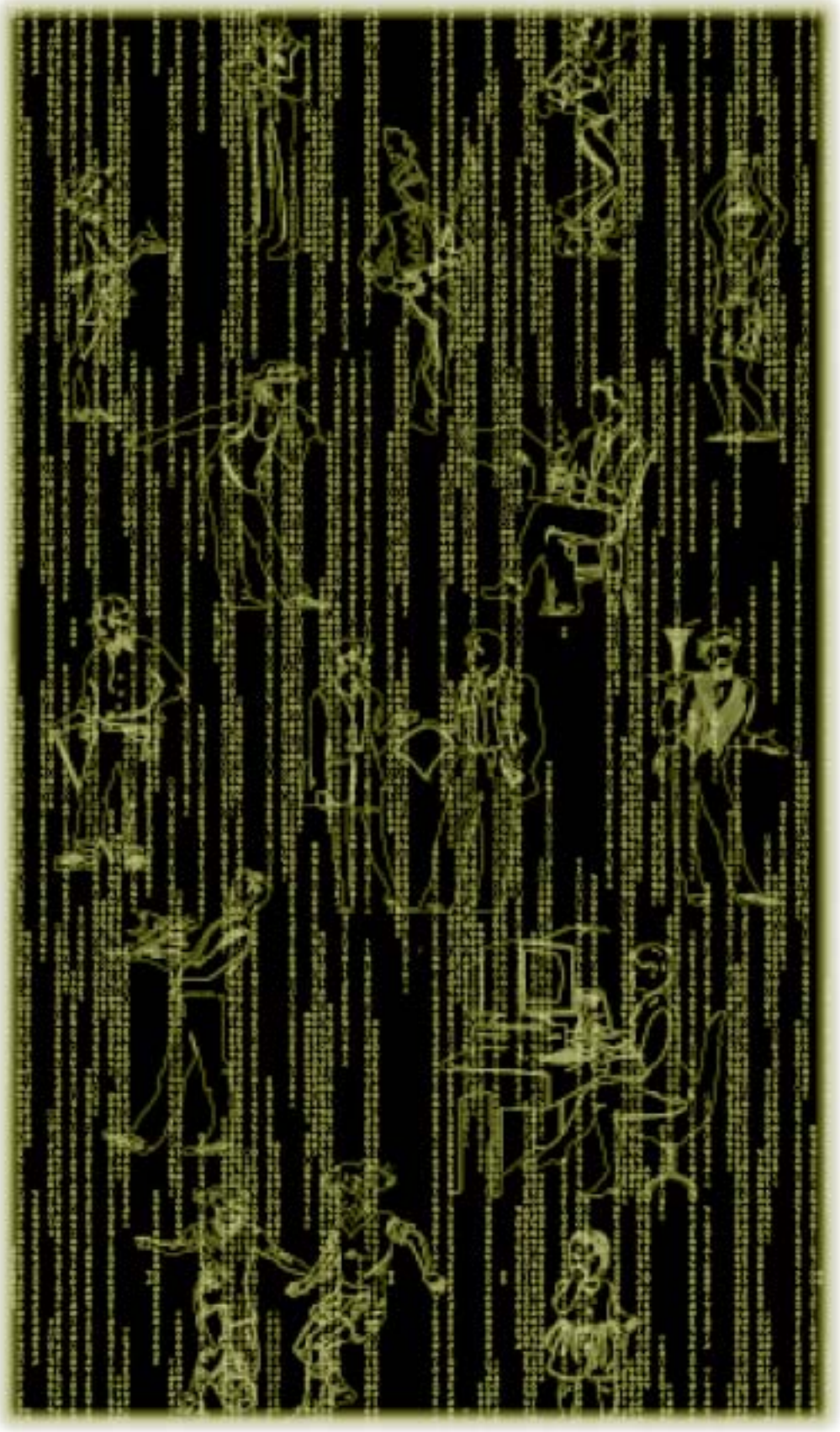
- Davidow, W.H., M.S. Malone. (1992). The Virtual Corporation, Harper
- Mowshowitz, A. (1994). Virtual Organization: a Vision of Management in the Information Age. The Information Society **10** (4):267-288
- Mowshowitz, A. (1997). Towards a General Theory for the Virtual Organization. Communications of the ACM **40** (9):30-37

4 http://www.maptive.com

4

Ethical and Legal Aspects

This part will focus on ethical and legal aspects of data mining. The ethical aspects are applicable in almost every situation, regardless of the specific country. Since we cannot cover the whole topic here, we have chosen to focus the ethics part on a current topic, web mining. You will notice, however, that many of the issues raised here are applicable to other situations as well (such as customer relations management). The chapter on ethical aspects is the Master thesis of Lita van Wel, realized for the STT data mining project. The legal part (Chapter 4.2) will provide a discussion from a point of view of of fair information practices, completed with views on the legitimacy of decision rules and on regulation of data mining for law enforcement.



4

4.1 Web mining in a Business Context: an Ethical Perspective

Lita van Wel[1]

4.1.1 Introduction

The World Wide Web can be seen as one of the largest databases in the world. This huge, and ever-growing, quantity of data is a fertile area for data mining research. In the introduction of this book, Meij describes data mining[2] as the process of extracting previously unknown information from (usually large quantities of) data, which can, in the right context, lead to knowledge. When data mining techniques[3] are applied to web data, we speak of web mining. In Chapter 5.6 (Web mining), Kosala, Blockeel and Neven explain that in the web mining context the term 'mining' is often used in a more general sense than just referring to data mining techniques in the classical sense. Therefore, they define web mining as "the whole of data mining and related techniques that are used to automatically discover and extract information from web documents and services"[4]. Note that again in the right context this can lead to the discovery of knowledge. This broad definition will be used in this section.

1 drs L. van Wel, LdeWit@chello.nl, Eindhoven University of Technology, Department of General Sciences, Eindhoven, The Netherlands. http://www.tm.tue.nl/aw/

2 In this study the term 'data mining' refers to the entire Knowledge Discovery in Databases process (KDD). Within KDD terms, data mining is just one step in the entire process. However the term 'data mining' is often used to refer to the entire process. And as there is no common use of a term like 'Knowledge Discovery in the Web' (as a special database), we shall use the more commonly used terms 'data mining' and 'web mining'.

3 More detailed descriptions on data mining techniques can be found in other parts of this book, or in other publications like [Fayyad, 1996a] in which the KDD field is explored. Their view is also presented in a paper [Fayyad, 1996b].

4 See Chapter 5.6, Web mining.

By looking at web mining from an ethical perspective, we shall discover a field of tension, between advantages on the one hand and disadvantages on the other. As ethics is the branch of philosophy concerned with the nature of morals and moral evaluation [Vos, 1995], an ethical perspective will raise questions such as what is right or wrong, what is beneficial or harmful. Ethical research focuses on three types of problem. First, there are situations in which normative principles are clearly disregarded. Secondly, there are ethical problems concerning new issues (types of problems that do not match existing cases) and the way in which traditional principles could be applied. The third type of ethical situation deals with normative conflicts. A normative conflict appears whenever there are both good and bad sides to a matter. The issue of web mining is a normative conflict where good refers to the benefits of web mining and bad to its possible harmful implications; in other words the ethical values that are threatened. Values are core beliefs or desires guiding or motivating attitudes and actions, and determining how people behave in certain situations. As ethics is a reflection on morality, ethical values could be described as that which subjects affirm as moral in human behavior [Xiaohe, 1998]. Thus ethical values have a normative function and are the motive for moral human behavior. A value can be seen as a global goal. Such a goal needs to be driven by a means, presented by more specific norms. For instance, the value of privacy is driven by norms such as respecting someone's private life and not misusing someone's personal data. Norms would be meaningless without values.
Knowledge discovered after mining the web, could pose a threat to people, when for instance personal data is misused. However, it is this same knowledge factor which can imply lots of different advantages, as it is of high value to all sorts of applications concerning planning and control. Kosala, Blockeel and Neven have already described some specific benefits of web mining, such as improving the intelligence of search engines. Web mining can also contribute to marketing intelligence by analyzing the web user's on-line behavior and turning this information into marketing knowledge.
This normative conflict, an area of tension between the advantages on the one hand and the disadvantages (threatened values) on the other, is the subject of this study.

It should be noted that ethical issues could arise from mining web data which do not involve personal data at all, such as data on different kinds of animals, or technical data on cars. However, this section is focuses on web mining that does in some way involve personal data. We shall only look at the possible ethical harm that can be done to people, which means that harm done to organizations, animals, or other subjects of any kind are not a part of the scope of this study. Since most web mining applications are currently found in the private sector, this will be our main focus. So web mining involving personal data will be

viewed from an ethical perspective in a business context. Note that this paper is not intended to be of a moralistic nature. Within the ethical perspective of this normative conflict, it is clearly recognized that web mining is a technique with a large number of good qualities and potential. However, examination of the possible threats might create certain awareness, leading to a well-considered application and further development of this technique.

This paper reads as follows. We will discuss the different ways to mine the web. To grasp the possible problems we will introduce two different categories of web mining. We will illustrate those categories by describing some fictitious cases. In Section 4.1.3 we will discuss the possible benefits of the different categories of web mining, using the cases to illustrate those benefits. Then, in Section 4.1.4, we describe the ways in which ethical values might be threatened by web mining. Again, the cases will be used to illustrate those ethical objections. Section 4.1.5 discusses the field of tension by analyzing some possible arguments debating the seriousness of the different ethical objections in the field of tension. The arguments are derived from interviews with web mining experts and a desk study. In our attempt to map the field of tension, we propose some solutions to the identified problems in Section 4.1.6. The ethical context presented here might help to find suitable solutions, so that web mining can be both as profitable and as harmless as possible. We will end with some concluding remarks.

4.1.2 Categories of web mining

There are different ways to mine the web. To structurally analyze the field of tension we need to be able to distinguish between those different forms of web mining. The different ways of mining the web are closely related to the different types of web data. We can distinguish actual data on web pages, web structure data regarding the hyperlink structure within and across web documents, and web log data regarding the users who browsed the web pages.
Therefore, in accordance with [Madria, 1999][5], we shall divide web mining into three categories. First, there is *content mining* to analyze the content data available in web documents. This can include images, audio files etc., however, in this study content mining will only refer to mining text. Secondly, there is the category of *structure mining* which focuses on link information. It aims to analyze the way in which different web documents are linked together. The third category is called *usage mining*. Usage mining analyses the transaction data that is logged when users interact with the web. Usage mining is sometimes referred to as 'log mining', because it involves mining the web-server logs. Structure mining is often more valuable when it is combined with content min-

5 See also the contribution of Kosala, Blockeel and Neven to this book.

ing of some kind to interpret the hyperlinks' contents. As content and structure mining also share most of the same advantages and disadvantages (as we shall see later on), we shall discuss them together, considering them as one category. It should however be noted that content and structure mining are not the only mining types that can be combined in one tool. Mining content data on a web site can for instance be of added value to usage mining analyses as well [6]. The combination of the different categories of web mining in one tool, could increase the value of the results. Web usage mining, however, is quite distinct in its application. As it is also used for different advantages and threatens values in a different way, we shall discuss it separately.

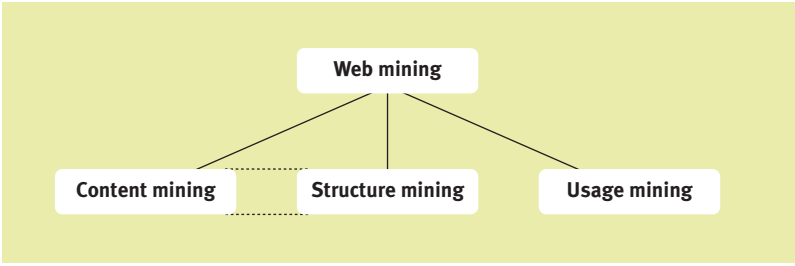


Figure 1
Three different categories of web mining.

The two remaining categories will be used to see whether the different kinds of web mining will lead to different beneficial or harmful situations. But first the categories will be illustrated by the following cases.

Content and structure mining

Sharon really likes to surf on the web and she loves to read books. On her personal homepage she likes to share information about her hobbies (surfing and reading) and she mentions her membership of a Christian Youth Association. In her list of recommended links she also refers to the web site of the Christian Youth Association. She has included her e-mail address in case someone wants to comment on her homepage.

An on-line bookstore decides to use a web mining tool to search the web for personal homepages to identify potential clients. It matches the data provided on homepages to existing customer profiles. After analyzing the content and the structure of the mined pages, they discover that people who link to Christian web sites of some kind all show a great interest in reading and generally spend a lot of money on buying books. So if the bookstore then makes a special effort to solicit Christians as customers, it might lead to more buying customers and an increase in profits. The web mining tool, not only provides the bookstore with a list of names, but it also clusters people with the same interests and so on. After analyzing the results, Sharon is identified as a potential, high-consuming customer. The bookstore decides to send Sharon a special offer by e-mail.

6 Notice, however, this mostly refers to content mining at a smaller scale (one web site). Read more about this in a paper like 'Integrating web usage and content mining for more effective personalization' by [Mobasher, 2000].

Sharon is somewhat surprised at receiving the e-mail from this bookstore. She has never heard of the store before and she wonders how they could have obtained her e-mail address. A bit annoyed, Sharon deletes the e-mail hoping she will never be bothered by this bookstore again.

Usage mining

Sharon always goes to her 'own' on-line bookstore. She frequently visits its web site to read about the newest publications and to see if there are any interesting special offers. The on-line bookstore analyses its web server logs and notices the frequent visits of Sharon. By analyzing her click streams and matching her on-line behavior with profiles of other customers, it is possible to predict whether or not Sharon might be interested in buying certain books, and how much money she is likely to spend on that. Based on their analyses, they decide to make sure that a banner is displayed on her browsing window that refers to a special offer on a newly published book that will most likely be of interest to Sharon. She is indeed appealed by the banner and she follows the hyperlink by clicking on it. She decides to accept the special offer and she clicks on the order button. On the on-line ordering form there are a lot of fields to be filled in, some don't really seem to be relevant, but Sharon does not see any harm in providing the information that is asked for. The people at the bookstore who developed the ordering form intend to use the data for marketing intelligence analyses. In the privacy statement that can be found on the bookstore's web site this intended use of the collected information is explained. The statement also contains a declaration that the gathered information will not be shared with third parties.
However, after a while the on-line bookstore discovers that web users from a certain provider hardly ever buy anything, but do cause a lot of traffic load on their server. They decide to close the adaptive part of their web site to visitors who use that certain provider. Sharon happens to be one of them and the banner in her browser window no longer displays special offers, when she visits the site of the bookstore.

4.1.3 Advantages of web mining

Web mining is attractive for companies because it brings several advantages. In the most general sense it can contribute to the increase of profits, be it by actually selling more products or services, or by minimizing the costs. In order to do so, marketing intelligence is required. This intelligence can focus on marketing strategies and competitive analyses, or on the relationship with the customers. The different kinds of web data that are somehow related to customers will then be categorized and clustered to build detailed customer profiles. This will not only help companies to retain current customers by being able to provide more

personalized services, but it also contributes in the search for potential customers. The example case clearly illustrates this.

By analyzing the web log data (usage mining), Sharon's favorite bookstores discovered that Sharon is a potential buyer. They were able to make her a tempting offer by displaying a specific banner on her browser window. The other bookstore was able to identify Sharon as a potential customer by searching the web for homepages and analyzing the data on the pages (content and structure mining). She was sent a special offer by e-mail, which would most probably match her preferences.
From Sharon's point of view we could say that she was pleased by the fact that the web site of her favorite bookstore displayed an interesting banner and she was not aware of any missed offers from this bookstore. And although in this scenario Sharon was a bit annoyed by the unsolicited e-mail send by the first bookstore, she might just as well have been attracted by the offer and she might even have become a customer of this bookstore.

Clearly both web mining categories contribute to the general goal of gaining marketing intelligence, be it each in its own way.

Content and structure mining

One of the things that make the web so special is that its content and structure data are largely publicly available[7]. So in theory everybody can perform content and structure mining on the web, provided they have the right knowledge and sufficient means for it. Content and structure mining tools are developed by organizations that specialize in web search and data mining technologies. Techniques such as finding related words based on frequent occurrence within the same page, result in a larger number of interesting patterns. Applying data mining techniques to web content data, could therefore result in better ways of finding relevant information on the web [Kosala, 2000]. Structure mining can aid to this goal, by identifying popular sites (so-called 'authorities'), for example, by analyzing the number of links that refer to a particular site [Madria, 1999]. Web content and structure mining are not only used to improve the quality of public search engines. Special search services can also be offered. Content and structure mining tools can for instance track down online misuse of brands[8], or analyze the content and structure of competitive web sites in detail to gain some strategic advantage[9]. With content and structure mining tools, things such as online curriculum vitae or personal homepages can be collected. After interpreting the personal data found on personal pages this information could be used for marketing purposes. Profiles on potential customers can be produced and more detailed information is added to profiles of current customers. So mining the web not only contributes to acquiring new customers, it can also aid in retaining existing ones.

7 [Custers, 2001] notes that recently there is a strong tendency for quid pro quo; companies that are only willing to let you enter their web site if you fill in an inquiry first, which mainly consists of questions for personal data. More about this in the next section.

8 See for instance a service offered by Cyveillance: http://www.cyveillance.com/

9 See for instance a service offered by WiseGuys: http://www.bluescope.nl/

Usage mining

Just like content and structure mining, usage mining also provides marketing intelligence [Büchner, 1999]. In contrast to content and structure data, web usage data is, however, not publicly available. Only those who provide the user access to the web and those who own the sites visited by the user are able to produce transaction logs. The advantages of usage mining therefore lie in building profiles based on these transactional data. Web logs provide an exciting new way of collecting information on visitors in a way that allows a site owner to actually see what the visitor is looking at [Khabaza, 2000]. Any action that tailors the web experience to a particular user, or set of users, can be seen as 'web personalization' [Mobasher, 2000]. Web personalization is often regarded to be an indispensable part of e-commerce. The ability to track web users' browsing behavior down to individual mouse clicks makes it possible to personalize services for individual customers on a massive scale [Srivastava, 2000]. This 'mass customization' of services not only helps customers by satisfying their needs, but also results in customer loyalty. Due to a more personalized and customer-centered approach, the content and structure of a web site can be evaluated[10] and adapted to the customer's preferences and the right offers can be made to the right customer. When web sites automatically improve their organization and presentation by learning from visitor's access patterns, we speak of adaptive web sites[11]. Web usage mining is also used to create personalized search engines, which can understand a person's search queries in a personal way by analyzing and profiling the user's search behavior[12]. It offers better, more personalized information after filtering out the links which are unlikely to interest a user. Thus, mining web usage data can improve personalization, create selling opportunities and, ultimately, lead to more profitable relationships with customers.

Let us now have a look at the way in which individual web users can benefit from all these advantages. Clearly, web users benefit from web mining when the techniques are used to improve the quality of public or personalized search engines. It assists them while navigating through the huge and ever- growing web. When companies provide more personalized services on their (adaptive) web sites, the individual web user can benefit from offers and services that are adjusted to his personal needs and preferences. Some even argue that the growing accuracy of profiles will lead to less unsolicited marketing approaches, for when a company is approaching a web user, it is most likely to be for something he or she actually is interested in.

Apparently a lot of the advantages of web mining are based on customer profiling. It is often more cost efficient to look at a group instead of looking at each individual, because groups are cheaper and easier to approach (for instance by

10 [Spiliopoulou, 2000] explains in more detail how web logs can be analyzed and used for site evaluation.

11 Read more on adaptive web sites in Section 5.6.3, Mining for adaptive web sites.

12 For instance Subme, an experimental search engine that uses artificial intelligence to understand your search queries in a personal way (http://www.subme.com), developed by the Dutch company 'SmartHaven'.

placing an ad in the right magazine instead of mailing every individual member of the group) [Custers, 2001]. These so-called group profiles can also be of added value to individual profiles, because of the fact that some individual characteristics only emerge after looking at the individual from a group perspective [Custers, 2001]. A person who never reads may still be interested in books, because he lives in a neighborhood where everybody reads a lot and he likes giving neighbors books on their birthdays. An individual profile will not show this characteristic, while the neighborhood group profile does. Furthermore, group profiling can result in better predictive values, making it easier to point out potential customers.

content and structure mining	**usage mining**
business level	*business level*
create more detailed (potential) customer profiles	create online customer profiles
direct marketing	direct (online) marketing
improved web filters	increase the value of each visitor
find related words by their frequent occurrence in the same page	mass customization
	adjust web filters to individual preferences
learn which topics are related to each other by occurrence of links between topics	register what visitor looks at
	create adaptive web sites
identify web sites that are of high general interest ('authorities')	improve content and structure of web sites
individual level	*individual level*
improved search services	personalized search engines
improved web filters	personalized web filters
fewer unwanted marketing approaches	adaptive web sites
	fewer unwanted marketing approaches

Table 1
Some advantages of web mining.

Looking at Table 1, it can be seen that web mining can be beneficial to both businesses and individuals. However, to make sure that this technique will be further developed in a properly thought-out way, we will also have to consider its possible disadvantages. Awareness of all the possible dangers is of great importance for a well-guided development.

4.1.4 Values threatened by web mining

In this section we point out that web mining, involving the use of personal data of some kind, can lead to the disruption of some important normative principles. One of the most obvious ethical objections lies in the possible violation of peoples' privacy. However, when the judgment and treatment of people is based on patterns resulting from web mining, the value of individuality is also threatened. Before discussing this value, we will first have a look at the issue of privacy violation.

Privacy

In the case of Sharon we see that in a way her privacy was threatened by both bookstores. The first bookstore decided to send Sharon a special offer by e-mail after retrieving her e-mail address from her personal homepage (content mining). And her favorite bookstore analyzed her browsing behavior (usage mining) and matched that data with a profile on browsing behavior of their existing customers to be able to tell whether Sharon would be likely to actually buy something and how much she would be most likely to spend on such a purchase. They did this without her knowledge and consent, but there was no real personal data (in the traditional sense) involved, so one could argue that her privacy has not been violated. All they did was make sure that a specific banner would be displayed on her browsing window the next time she visited their web site. The way to do this, however, is to place a small file on Sharon's computer that will make sure that her computer is recognized the next time she visits the web site. That in itself could be seen as a violation of Sharon's privacy. Sharon did not experience it as a violation, because she was probably not even aware of it and she was pleased with the special offer. Later on the bookstore again violated Sharon's privacy without her knowing it, for she was not aware of the fact that the web site owner analyzed all usage data and concluded that users from a certain provider would no longer receive special offers.

Privacy is generally defined as the quality or condition of being secluded from the presence or view of others. Privacy can involve one's private life, referred to as relational privacy, but it can also involve one's personal data. The latter is referred to as "informational privacy, which can be defined as an individual's ability to autonomously control the unveiling and dissemination of data referring to his private life" [Vedder, 1998, page 115]. Privacy issues due to web mining often fall within this category of informational privacy [Vedder, 1998; Tavani, 1999a]. Therefore this study will be focusing on this category. In the remainder of this paper the term privacy will be used as referring to informational privacy. There are some differences between privacy issues related to traditional information retrieval techniques and the ones resulting from data mining. The tech-

nical distinction between data mining and traditional information retrieval techniques does have consequences for the privacy problems evolving from the application of such techniques [Tavani, 1999a]. While in traditional information retrieval techniques one has to 'talk' to a database by specifically querying for information, data mining makes it possible to 'listen' to a database [Holsheimer, 1999]. A system of algorithms searches the database for relevant patterns by formulating thousands of hypotheses on its own. In that way interesting patterns can be discovered in huge quantities of data. [Tavani, 1999a] argues that it is this very nature of data mining techniques that conflicts with some of the current privacy guidelines formulated by the OECD [13]. These guidelines correspond to the European Directive 95/46/EC of the European Parliament and the Council of 24 October 1995. Every European Union member state has to implement these basic guidelines in their national laws [14]. It is obviousthat the guidelines are highly influential, therefore the definitions and principles should not be underestimated. Principles such as the 'use limitation principle' and 'the purpose specification principle' state that information can not be used for other purposes than the one it was originally retrieved for, and that this purpose has to be clearly explained to the person whose data is being mined before the actual collection. However, one of the features of data mining is the fact that it is not clear what kind of patterns will be revealed. That makes it impossible to clearly specify the exact purpose and notify the data subjects in advance [Tavani, 1999a] [15]. Besides, data mining is often performed on historical data sets, which also makes it rather difficult to comply with these guidelines. With web mining, and the way in which the different types of web data can be collected and analyzed by data mining tools, it is also difficult to comply with the current guidelines.

Content and structure mining

Apart from the general problem of web mining not complying with the current guidelines, another problem arises: content and structure data are publicly available. People might have placed certain bits of information on their homepage for certain purposes and within a certain context. However, when web data is mined, it can be used for totally different purposes, taking it completely out of context. Because most of the web data has been made public by a web-user's own leave, it is debatable whether this kind of information deserves to be protected at all. However, it can be argued that it is wrong to assume that an aggregation of information does not violate privacy if its parts, taken individually, do not [Nissenbaum, 1997]. A single fact about someone can take on a new dimension, when it is combined with another fact, be it about the same individual or about others. Although different public facts might not be considered to be harmful, the association between those facts can violate someone's privacy [Fulda, 1999].

13 In 1980 the OECD (Organization for Economic Cooperation Development) formulated some internationally accepted principles regarding the collection, use and unveiling of personal data. An online description can be found on: http://www.oecd.org/dsti/sti/it/secur/prod/PRIV-EN.HTM

14 In The Netherlands, the Directive's guidelines are implemented in the 'wet Bescherming Persoonsgegevens' (WBP), which will replace the current 'wet Persoonsregistraties' (WPR).

15 Nevertheless, usually it is possible to formulate the purposes at a higher level of abstraction (Ed.)

Usage mining

With usage mining, the privacy in public discussion takes on an extra dimension. Web log data are not publicly available, yet the data do represent someone's actions within a public environment. Let us compare it to a shopping street. Web users are the people that move around in this shopping street. When entering a store, a person will still find himself in a public area, because everybody can freely enter the store. The area is, however, confined to a single store instead of an entire shopping street, so only people who are inside the store can see and watch this person. Let us imagine that there is nobody else in the store, except for the people who work there. Those people are the only ones that can observe him. A web site could be seen as such a store, with only one visitor and the personnel inside. Once a web user enters a web site, the people who manage and own the site are able to observe his steps. When there are video cameras in every corner of the store, a visitor loses his privacy. As the store only uses them to protect itself against burglars, however, this is usually not considered to be a violation. Consider the possibility of the store using the camera to record precisely what products each customer looks at and for how long. They could also decide to try and relate that to some more personal details they have recorded, such as the gender of the visitor, the clothes he is wearing, his kind of haircut, or the color of his skin. Once a person is on a web site, certain data are logged, but not the kind of personal data that can be recorded with a video camera. Web log data do not actually identify a person, but they do identify a web user; a user who has characteristics such as a certain IP-address, date and time of entering and leaving the site, path of followed hyperlinks (click stream), type of browser used, and so on. So they do not know his name or what he looks like, but the next time he visits the web site, he will be recognized as a regular visitor by use of cookies. Cookies are small files that are placed on the hard disk of the web user during his browser session. The cookie will make sure that the web user's computer will be recognized and identified the next time the same web site is visited. Therefore cookies can be used to track a user's online movements and enable the creation of a profile. Some site owners allow advertisers to place banners referring to their own (ad)server on a web site. By loading the banner ad, a cookie is placed on the web user's computer. That way, online advertising companies are also able to track (part of) a user's movements on the web.

ISP's[16] are the only ones that can directly link all web surfers' behavior to their personal data. When a web user starts his surf session, an ISP knows who this user is, because they are the ones that provide him access. In order to obtain an access account, the web user had to provide the ISP with some of his personal data[17]. Every time he uses the web, he browses via the ISP's server, so they can monitor all the moves he makes. The Dutch law relating to lawful interception obliges ISP's to keep logs of all those user transactions in case the government

16 Internet Access Provider (IAP) is an organization that provides Internet Access (e.g. PPP, SLIP, shell, or UUCP accounts) and possibly other services. Internet Service Provider (ISP) is an organization that provides one or more basic Internet Services such as Internet access, web site hosting, or DNS support for domains. In casual conversation, IAP and ISP are interchangeable terms. As the term ISP is more commonly used, we shall speak of ISP's, even though in a strict sense we actually refer to IAP's.

17 This is the common situation, however some providers allow anonymous accounts.

might need it for criminal investigations [Artz, 2000]. However, law also restricts ISP's. In the Netherlands, ISP's are legally regarded as telecommunications service providers and they have to adhere to the new Dutch Telecommunications Law[18] (since 1998). This law stipulates that telecommunication service providers are obliged to erase, or make all traffic data relating to subscribers upon termination of the call anonymous. The only exceptions to this rule are for traffic data that are necessary for the purpose of subscriber billing, marketing purposes (provided that the subscriber has given his consent), control of telecommunications traffic, supply of information to the customer, the settling of billing disputes, detection of fraud, or otherwise allowed (or even obliged) by legislation. Notice, that mining for marketing purposes is only allowed with the customer's consent. An ISP can specifically ask for this consent when a new customer subscribes. However, not all users are able to foresee all consequences, when agreeing to the use of their data for marketing goals. Moreover, as explained previously, it is difficult to clearly state the purpose of web mining analyses due to its exploratory nature. This could potentially lead to situations of privacy-violation. However, most Dutch Internet providers are member of the NLIP (Organization of Dutch Internet Providers[19]), which binds them to privacy guidelines as part of the code of conduct that every member has to honor. The NLIP web site states that about 80% of Dutch people who go on-line use a provider that is a member of the NLIP. So ISP's can indeed connect personal data to their web log data, but they are limited in their freedom to analyze the data and act on it. Besides, once a web user fills in a form on a web site, the site owner also possesses the personal data and can link it to the web log data of his web site.

The latter refers to another type of web usage mining: the use of all kinds of forms on the web. Lately there has been a tendency to trade information as a quid pro quo [Custers, 2000]. Web users often have to fill out an inquiry simply to gain access to a web site. Or there are fields to be filled in on on-line ordering forms that are of no relevance to the purchase. When a user fills in an on-line form of any kind, the data he shares can be used for customer profiling. By sending back the form, the web log also registers the IP address of the web user and his personal data can therefore be linked to his browsing behavior on that particular web site. Although a user decides for himself whether or not he will fill in a form, the way the data is used after collection might still violate his privacy. This is especially the case, when he is not aware of the fact that his personal data is being classified and clustered into profiles. Moreover, it is often unclear to a web user how some apparently trivial piece of data might result in non-trivial patterns.

A shortcoming of the traditional concept of informational privacy lies within the assumption that personal data concerning privacy, originally consist of identi-

18 In 1998 the Dutch Telecommunications Act came into force. The rules can be found on http://www.minvenw.nl/dgtp/home/beleid/juridisch/m_juridisch.htm. The text is also available in English: http://www.minvenw.nl/dgtp/home/wetsite/engels/index.html

19 The 'Vereniging van Nederlandse Internet Providers' looks after the interests of Dutch Internet Providers (http://www.nlip.nl/).

fiers of individual persons and that the data continue to contain those identifiers [Vedder, 1999]. This makes it difficult to discuss the problematic situations in which data are abstracted from personal data and used for production and application of group profiles and generalizations within the concept of privacy[20]. However, the value for people to be judged and treated as individuals is threatened.

Individuality

The special offers made, or *not* made, to Sharon are based on profiles produced by the bookstores. They both matched her personal characteristics (be it online behavior, or personal data collected from her homepage) to other profiles, to predict what properties might be assigned to her. The way they actually judge and treat her (whether or not to make her a special offer, and if so what kind of offer) depends on the group characteristics of the profiles Sharon's personal profile is matched with. One of the bookstores mined the web to find homepages and then, after mining the content and structure of those homepages, discovered that people who link to Christian web sites of some kind all show great interest in reading and generally spend a lot of money on buying books. So, if the bookstore were then to make a special effort to solicit Christians as customers, it might be able to get more well-buying customers and increase its profits. Note that it is possible that Sharon does not actually manifest every group characteristic as an individual characteristic. For instance, Sharon is being denied special offers, simply because of the fact that she uses a certain Internet provider (web log data), while she would definitely be interested in buying books. The decision not to make special offers to visitors who use a certain Internet provider is based on a generalization, and Sharon is one of the exceptions.

Individuality can be described as the quality of being an individual; a human being regarded as a unique personality. Individuality is one of the strongest Western values. In most Western countries people share a core set of values maintaining that it is good to be an individual and to express oneself as that individual. Profiling through web mining can, however, lead to de-individualization, which is defined as "a tendency to judge and treat people on the basis of group characteristics instead of on their own individual characteristics and merits" [Vedder, 1999, page 275].

Vedder distinguishes between two types of group profiles[21]. Distributive group profiles are profiles in which every group characteristic is actually shared by every individual member of the group. However, in non-distributive group profiles, not every characteristic of the group is shared by every individual member. The properties in non-distributive profiles apply to individuals as members of the group, but the individuals themselves need not in reality exhibit all of these properties. In non-distributive group profiles, data are framed in terms of proba-

20 [Vedder, 1998; 1999; 2000] suggests a broader definition of privacy (referred to as categorical privacy), which would allow group characteristics that are applied as if it were individual characteristics, to be seen as personal data. However, [Tavani, 1999b] beliefs that such a suggestion would not be suitable, for it might lead to the need for new privacy categories with the introduction of every future technology.

21 Vedder refers to this distinction in all of his papers listed in the reference list at the end of this paper.

bilities and averages and so on, and the data are therefore generally made anonymous. When data is made anonymous before producing a profile, the discovered information no longer links to individual persons and there is no direct sense of privacy violation. The profiles do not contain 'real' personal data, which is commonly defined as data relating to an identified or identifiable person. Basic principles from the European Directive such as the 'collection limitation principle' and the 'openness principle' state that personal data should be obtained by fair means, with the knowledge or consent of the data subject, and that a subject has the right to know what personal data are being stored. The 'individual participation principle' even gives a subject the right to object to the processing of his data. All of those principles depend heavily on the assumption that there is some kind of direct connection between a person and his or her data. However, in anonymous profiles this direct connection has been erased. Nevertheless, the profiles are often used as if they were personal data. This sometimes makes an impact on individual persons stronger than that made by the use of 'real' personal data. The information cannot be traced back to individual persons. Therefore, individuals can no longer protect themselves with traditional privacy rules. However, when group profiles are used as a basis for decision-making and formulating policies, or if profiles somehow become public knowledge, people will be judged and treated as group members rather than unique individuals. By threatening the value of individuality, people could even be discriminated [22]. This is for instance the case when profiles that contain data of a sensitive nature are used for selections in certain allocation procedures. Some criteria (usually) are inappropriate and discriminatory to use in decision-making, such as race, religion, and so on. This is especially true when the information is irrelevant (and often illegal to use) for decisions, such as turning someone down for a job, or not giving him a loan and so on [Johnson, 2001].

As both categories of web mining are used for profiling, both categories pose the same threat to the value of individuality and the closely related value of non-discrimination. Web mining ultimately jeopardizes the values of fair judgment and fair treatment. In the next section we shall try to demonstrate the seriousness of these dangers.

22 Discrimination can be defined as the act of making an invalid, unfair, or hurtful differentiation. It refers to any situation in which a group or individual is treated differently based on something other than individual reason, usually their membership in a socially distinct group or category. Discrimination can be viewed as favorable or unfavorable, depending on whether a person receives favors or opportunities, or is denied them.

4.1.5 The field of tension

The mere fact that it is possible to describe both advantages and disadvantages of web mining proves that there is a field of tension around the development and application of web mining techniques. Still the question remains: how much tension is there in this field? If the disadvantages were to totally outweigh any possible advantage, or to put it the other way around, if the advantages were to clearly outweigh any disadvantage, there would hardly be any tension at all. In the case of web mining, however, there clearly is a certain amount of tension. All the benefits obviously show that web mining is a highly valuable technique, which is being developed and applied on a large and growing scale. However, the threats to some important values tend to be rather serious and will create tension in the web mining field. In order to 'measure the amount of tension', we shall try to determine the seriousness of the dangers. We shall do this by discussing the possible arguments or views of people that do not foresee any danger in web mining with regard to ethical values. To gain some insight into current web mining practices and the attitude of web miners to ethical issues involved, six people who apply web mining in a business context, were interviewed. These interviews were based on eight general questions, which are included in the appendix of this paper. The list of questions was mainly used as a checklist to make sure that all the topics of interest would be discussed. The interviews combined with a desk study[23] teach us that people prefer to focus on the advantages of web mining instead of discussing the possible dangers. Moreover, it revealed several different arguments to support the view that web mining does not really pose a threat to the ethical values described. After combining some of the arguments, it could be condensed to a list of seven main arguments, as shown in Table 2.

Table 2
Possible arguments against the danger of web mining.

	possible arguments
1	Web mining itself does not give rise to new ethical issues.
2	There are laws to protect private information and on-line privacy statements guarantee privacy.
3	Many individuals have simply chosen to give up part of their privacy.
4	Most data collected is not of a personal nature or used for anonymous profiles.
5	Web mining leads to fewer unsolicited marketing approaches.
6	Personalization leads to individualization instead of de-individualization.
7	Most customers like to be recognized and treated as special customers, so it is not considered a violation of privacy to analyze usage interaction.
8	Public data on the web is there for the taking.

23 The interviews and desk study are part of my current Master's Thesis Research for the Technology and Society program at the Technical University Eindhoven, The Netherlands.

If all, or most of these arguments can be refuted, we would have to conclude that there definitely is a substantial amount of tension in the web mining field.

1st argument

Web mining itself does not give rise to new ethical issues.

Indeed, most of the possible dangers come from group profiling (in particular non-distributive group profiles), and since group profiling was being done before data mining techniques were known, the issues could be considered old news. It is however important to realize that data mining does significantly enlarge the scale on which profiling can be practiced. A lot more data can be collected and analyzed, and (as explained) new patterns can be found without asking for it in specific queries and hypotheses. Using these data mining techniques on web data creates another level of decision-making, in which companies can use large amounts of detailed profiles based on the behavior and characteristics of web users. So although the ethical issues are not actually new, they are lifted to another dimension. This requires a new perspective on both ethical values and the development and application of web mining.

2nd argument

There are laws to protect private information. Besides, privacy policies found on many web sites guarantee privacy, so why worry?

The law is not, and never will be fully sufficient. We have already discussed some of the existing shortcomings of the law with respect to the privacy problems. For instance, the fact that current privacy laws only offer protection for the misuse of identifiable personal data, shows us there is no legal protection for the misuse of anonymized data used as if it were personal data. It should be noted that not everybody agrees on this point. For instance [Schreuders, 2001] agrees that group profiles are not considered to be personal data (within Dutch law), but he claims that the application of group profiles is considered to be "processing of personal data" and is therefore protected by the Dutch WBP[24]. Although this debate on whether or not group data are protected to the same degree as personal data has not yet been settled, it is clear that law does not fully grasp the ethical issues concerning web mining. It is more effective to solve a problem in society without, or before having to create laws against it. Legal measures are no guarantee that no harm will be done. In other words, laws do not abolish the problems. Even if there are laws, general public support is needed to make them effective. Such a basis of public support can also stimulate self-regulation.

The growing number of on-line privacy policies is an example of self-regulating efforts. However, such policies are not found on every site, so there still will be a

24 [Borking, 1998] has also pointed out that Dutch privacy protecting rules only offer protection if personal data can be traced back to an individual. They state that, once (anonymized) group profiles are projected on an individual and combined with his personal data, privacy rules do apply, because this concerns the processing of personal data.

lot of sites which a person who is concerned about his on-line privacy should not visit. Besides that, it is not easy for a web user to search for and thoroughly read the privacy statements on every site he visits and to check whether the policy has changed every next time he visits the same site. Furthermore, the policies are often unclear about certain points. Take for instance lines such as 'will not share information with third parties'. Privacy declarations promising not to sell or give information to third parties are often not clear about who those third parties are, and more importantly, who are not. Stakeholders and co-operating businesses should also be regarded as third parties, but often are not. Major businesses that consist of several kinds of different smaller companies could exchange large quantities of private information, if they do not regard every sub-company as a third party. Thus in large companies information could be shared on a wide level; a lot wider than most people would foresee. A lot of other statements are often unclear as well. Take for instance lines that describe the use of collected data. If such a goal is just marked as 'marketing goals' then how can a user know what specific kind of marketing to expect? Privacy policies are often difficult to understand, hard to find, take a long time to read and can be changed without notice [Scribbins, 2001]. Apparently, regulation and self-regulation efforts do not offer sufficient protection for web users' privacy.

3rd argument

Many individuals simply choose to give up part of their privacy.

There is some truth to this argument. There are of course people who are willing to share their personal data with everybody, without wanting anything in return. They simply do not care about their privacy. Other people are only willing to share their personal data, if they get a product or a service in return. Those people might be tempted by web offers like SportsLine.com [25], who offer people saving points in return for personal data. They could also be willing to enter web sites that use a log in, where one can only enter the site, if he first subscribes and gives (some of) his personal data. Surely there also are people who will, under no circumstances, share their private information with organizations on the web. For most people, however, it depends on the situation whether or not they are willing to lose some privacy. However, one should notice that the price of privacy on the web has become rather high [Johnson, 2001]. Of course you can choose not to have an account with an Internet provider or – having such an account – to choose never to access information on the web, so that there are no records of what has been viewed. These choices reduce the quantity of information organizations on the web have about this person. However, he will have to give up a great deal. Not using the web does not seem to be a fair option; it would be a high price to pay for his privacy. Besides, most people using the web are not aware of the ways in which their web data can be analyzed. Is it fair to

25 SportsLine.com offers site visitors the option to sign in and earn points by surfing on the site. The points can be redeemed for rewards (http://ww1.sportsline.com/u/rewards/sportsclub.cgi).

say that people choose to give up their privacy, when they are not fully aware of the consequences of their actions? Can we expect people to be fully aware of those consequences? There are of course situations in which people are made aware and can then make informed choices about whether or not to give up their privacy in certain situations. In the next section we shall discuss some possible solutions.
But even though people might have some control over whether or not to give up their privacy in certain situations, it is rather difficult, if not impossible, for an individual to protect himself against the threat of de-individualization due to the use of non-distributive group profiles. Even if an individual were able to refuse his data to be used for profiling, that would not prevent profiling itself, and profiles could still be projected on this individual. So he could still be judged and treated according to group characteristics. The nature of group profiles makes it (almost) irrelevant whether one individual refuses to cooperate. The larger the number of people that do participate the more futile it becomes for one individual to refuse.

4th argument

The data collected is not of a personal nature and most web mining applications result in anonymous profiles, so why should there be a (privacy) problem? An argument often heard is: 'Our software is used to identify 'crowd' behavior of visitors of web sites. Therefore, if we don't know/care who you are, how can we be invading your privacy?'

When group profiles are made anonymous, individuals can no longer protect themselves, because by current law these group data are not considered to be personal data. If, however, such non-distributive group profiles are projected on web users, they are used as if they were personal characteristics shared by every person regarded to be a member of such a profile. Therefore even anonymous profiles can be harmful, leading to de-individualization and possible discrimination. This projecting of group profiles on individuals is often done, because of the ultimate goal of analyzing crowd behavior on web sites: to create clusters and categories of site visitors according to the clusters for instance in order to personalize the web site. It is not necessary to actually know the person (by name and address and so on) as long as this particular web user can be identified as being part of a certain cluster.

5th argument

The data mining technique will provide more accurate and more detailed information, which can lead to better, fairer judgment. So web mining leads to less unwanted marketing approaches, therefore why would people complain?

[Putten, 1999] even foresees a possible occasional alliance between privacy protectors and data miners, because of their mutual goal: less unsolicited marketing contacts. Web mining could definitely be beneficial to individuals, if there were less unwanted marketing approaches as a result of better profiling. The reverse is, however equally possible, where web mining techniques might lead to more profiles then ever before, and the chance that people are part of a larger number of profiles grows. Perhaps the profiles will be more accurate, so people will only be approached by businesses that will most likely make them a special offer matching their preferences. If, however, the quantity of offers gets too large, people will be annoyed anyway. Besides, even if marketing approaches match their preferences, people still did not ask to be approached. And although web mining makes it possible to provide businesses with more accurate profiles, businesses would probably still approach as many people as possible, because the more people are approached, the higher the probability of reaching the right person(s). Consequently web mining could easily lead to more unsolicited marketing approaches. Furthermore [Clarke, 1994] wisely states that "at some point, selectivity in advertising crosses a boundary to become consumer manipulation".

6th argument

Web mining is often used for personalization. This leads to individualization instead of de-individualization.

One of the main goals of web mining, when used for e-commerce, is indeed personalization; adjusting web sites and services to the individual wishes of each visitor. So it does appear to promote individuality instead of threatening it. However, if the personalization is done by creating non-distributive group profiles, it leads to de-individualization, because individuals are no longer judged and treated on their individual characteristics and merits, but on group characteristics. As personalization is often based on profiles and generalizations, it threatens the value of individuality.
Of course, some people might prefer a personalized approach and are even willing to provide some personal data. However, people are often not offered a choice and in those cases there is a thin line between personalization and personal intrusion. Web users could find on-line solicitation from a web site a hindrance rather than helpful [Mulvenna, 2000].

7th argument

Most customers like to be recognized and treated as special customers, so it is not considered a violation of privacy to analyze usage interaction.

This argument refers back to the privacy in public discussion for web usage min-

ing, already mentioned in the former section. When the man behind the counter recognizes someone as a regular client and greets him in a friendly way, this loss of privacy is not considered to be a violation. An association has been made in the public domain, however, that knowledge in this situation is only in the heads of people. The data stored in a person's brain is limited to the capacity of the brain and is not accessible to others, unless the person decides to share it with someone. Data that is digitally stored does not have to be limited by the size of the database; the database can simply be upgraded. Digital data can be easily distributed and is multi-accessible (by anyone who wants to use it and is granted access to the database). Intelligent data analysis is possible and data can easily be connected and compared to other data stored in the same or in another database. More importantly, while many — if not most — of the associations between different pieces of data stored in someone's head are forgotten, data mining not only finds unexpected associations, these associations can also be stored for future use and reference. This is not necessarily morally wrong, but it is an area where more work is required to articulate what is acceptable and what not in usage mining [Weckert, 2000].
Furthermore, one should be aware of the limited data quality. The possible changes in IP addresses and the simple fact that even when it can be established that it is the same computer (by use of cookies), some other person might be using it, show that recognition is definitely not flawless.

8th argument

What can be wrong with collecting public data from the web, when it is there for the taking?

In Section 4.1.4 we have already mentioned the privacy in public discussion concerning web content (and structure) mining. Even if data is publicly available on the web (for instance on someone's homepage), it can still be morally wrong to mine it and use it for certain purposes without the publisher's knowledge and consent. Furthermore, when different bits of information are assembled, aggregated and intelligently analyzed, they can invade someone's privacy, even if the parts individually taken do not [Nissenbaum, 1997]. This is based on the assumption that a single fact can take on a new dimension, when it is combined with other facts, or when it is compared with similar facts. Even non-identifiable data can become identifiable when merged. With regard to web mining, certain bits of data, that are not considered to be privacy violating, can be collected and assembled with other data, leading to information that can be regarded as harmful. For instance when data is connected to external databases, such as demographic databases[26].
As in the case of usage data, one should also be aware of poor data quality of content data on the web. As there is no overall control on web content (and

26 Which is for instance done by Webminer, see: http://www.webminer.com

structure) data and there are no rules[27], anybody can place any information on the web. In other words, reliability of content data on the web is quite poor.

This is not intended to be an exhaustive list of all possible counter-arguments to the statement that web mining threatens some important values. And eminently not all arguments have been refuted; some have just been placed in a different perspective. However, it has been made clear that the dangers described in Section 4.1.4 do pose a serious threat. The more tension there is, the more difficult it will be to solve the problems. One thing is clear, even though web mining is still in an early stage of development, there is no way back. We are already in the middle of this tense field. It would not be an option to just declare that the disadvantages are so strong that web mining should be abolished, nor would it be an option to choose in favor of the advantages and just ignore all the disadvantages and pretend there are no dangers. The only option is to travel right through the web of tension. Clearly this is not an easy task, for there is no marked path. Notice that this is not a case of balancing the disadvantages and the advantages. We should strive to benefit from web mining as much as possible, while causing the least possible harm.

27 Except for existing laws for instance on the freedom of speech and publication.

28 Part of the 'Five Steps to Protecting your Privacy Online – A Customer Tip Sheet', appendix 6 of [Scribbens, 2001].

4.1.6 Possible solutions

There are means to release some of the tension in the ethical field surrounding web mining. We shall distinguish between solutions at an individual and at a collective level. Solutions at an individual level describe some actions an individual can take to protect himself against possible harm. The solutions at a collective level refer to things that could be done by society (government, businesses or other organizations) to prevent web mining from causing any harm. Let us first list the different possibilities before actually discussing their practical relevance and possible contribution in releasing some of the tension in the web mining field.

Individual level

Use Privacy Enhancing Technologies (PET's)

When you want to enhance your privacy while surfing on the web, you should use tools to make sure that your actions are not (directly) traceable by the site owners of the different web sites that you visit[28]. There are different PET- tools available to protect ones privacy on the web[29]. Most of them combine things like managing the acceptance or rejection of cookies[30] and ads[31] and providing on-line pseudonyms. A problem with tools that provide on-line pseudonyms is that although your on-line actions are not directly traceable to you in person,

29 The Electronic Privacy Center (EPIC) has published the 'EPIC Online Guide to Practical Privacy Tools' on the Web, which provides a list of privacy enhancing tools categorized by area of application. See http://www.epic.org/privacy/tools.html.

30 A cookie manager is included in most Internet browsers, see for instructions: http://www.junkbusters.com/ht/en/cookies.html#disable

31 By blocking ads, the ad on the server of the on-line advertising company is not able to register your visit to the page on which the ad was placed. Because ads are not loaded by your browser when visiting a site, the advertiser never gets to place a cookie on your computer. Another reason to use ad-blocking software is that it reduces the wait for a page to be loaded.

your actions can still be traced back to actions being from one person. Even though it is unknown who this person is, the data collected could still be valuable. Therefore, most tools for anonymous surfing include a cookie and an ad rejection option, which diminishes the possibilities of tracing a web user.

Be cautious when providing (personal) information on-line

Do not reveal personal information inadvertently [Scribbins, 2001]. Do not publish too much personal information on your homepage and if possible use false data to enjoy certain on-line services. Do not give out personally identifiable information too easily. You are not obligated to give out personal information about yourself simply because a site asks for it or even demands it. Of course, if you want to buy something, you will have to give accurate billing information, but if you are registering with a free site that is a little too nosy for you, you could decide to provide them with pseudonymous information. (To prevent bots from grabbing your email from your homepage, you could give it only in bitmap (graphics) format, which requires human interpretation. [Ed.])

Check privacy policies on web sites

Although we have already discussed important shortcomings of some current privacy policies, users are still advised to look for these policies and read them carefully [Scribbins, 2001]. An increasing number of web sites are providing privacy policies that detail the sites' information practices. When you visit a web site that has no privacy policy, you could decide not to visit the site again, or you could write and tell the company that you would like to see them post a policy.

Collective level

Further development of Privacy Enhancing Technologies

The PET's currently available for individual users, are developed by businesses or other organizations. An interesting development in this field is the development of privacy empowering web mining tools[32]. These tools should enhance user control over the use of personal information. They could for instance be implemented in a user's browser[33] and provide assurance to users that their privacy is protected without having to read each web site's privacy policy. In other words, privacy empowering tools provide an automated way for users to gain more control over the use of personal information on web sites they visit.

Publish privacy policy

Web site owners should formulate a privacy policy and publish it on an easily accessible place at the web site. The policy should clearly state the way in which web log data will be analyzed and used. There already are a lot of web sites with privacy policies[34], but they are not always easy to find and understand [Scribbins, 2001].

32 The Platform for Privacy Preferences Project (P3P), a tool developed by the World Wide Web Consortium (W3C), contributes to this goal. P3P offers a way for users to automate the acceptance or rejection of a web site's requests for information, based on preferences users can set from their browsers. More details: http://www.w3.org/P3P/

33 Microsoft Internet Explorer 6 will support privacy preferences (using P3P); users can choose their own level of on-line privacy. More details: http://msdn.microsoft.com/workshop/security/privacy/IE6PrivacyFeature.asp

34 For instance large IT companies like Microsoft (http://www.microsoft.com/info/privacy.htm), Amazon.com (http://www.amazon.com/exec/obidos/tg/browse/-/468496/002-3078300-3767210), or Yahoo (http://privacy.yahoo.com/privacy/us/)

Web quality seal

As people might still not trust a company after reading its privacy policy, companies can back up their privacy statement with a seal program. A web quality seal could help the web user to gain trust in on-line businesses by helping web users find reliable, trustworthy businesses on-line. Sites that have successfully met the guidelines and requirements are able to display some sort of quality seal to show site visitors that the site owners respect their privacy and are trustworthy. Often a validation option is built in, where a web user can check whether the seal is legitimately used, by clicking on the seal. The seal-providing-organizations usually keep lists of qualifying sites. These kinds of privacy initiatives often have commercial aims, for instance to accelerate growth in the industry by promoting consumer trust and confidence. Another motivation might be found in prohibiting government regulation in electronic commerce. There are several web quality seal initiatives[35], but given the nature of the World Wide Web an international standard would be preferable.

Monitoring web mining activities

There should be a non-partial organization that monitors web mining activities and informs the public of those practices. Although it is quite hard to actually monitor web mining activities, it would be a valuable contribution if some monitoring organization were able to create a bit more transparency on the Web. To some extent this monitoring is done by organizations that offer privacy seal programs. They check that their members are acting in accordance with the required guidelines. Another initiative could be found in the contribution of web users organized in some sort of worldwide watch team[36]. They could assist in the search for possible harmful web mining practices on the web.

Create awareness amongst web users and web data miners[37]

Web users who are informed of the possibilities and the dangers of mining the different types of data on the web, are able to make well-informed and well-considered choices. Notice that there is no uniform group of web users, they represent a variety of different social groups, such as children, youth, adults, elderly. Within those groups there can also be different subgroups of web users and so on. This means that one cannot really approach 'the web users'. It should also be noted that not only current web users should be informed, but also potential web users.

Web data miners should have a certain degree of moral sensitivity to the possible harms of mining different types of web data. The organizations that mine the web for patterns have a lot of knowledge of the possible productions and applications of profiles. If they were properly aware of all the dangers of web mining, there might be a good basis for public openness on the use of web data.

35 For instance TRUSTe, the Better Business Bureau Online (BBBOnline), the Japanese Privacy Mark and CPA Web Trust.

36 Webguardian (a public Internet monitor that watches and documents consistent consumer problems and complaints of the Internet community) has constructed a world wide watch team in which web users can participate in the monitoring of disturbing practices on the Internet.

37 One way to create more awareness and moral sensitivity is by an initiative like 'netiquette' (http://www.albion.com/netiquette/). This network etiquette lists the do's and don'ts of on-line communication and covers both common courtesy on-line and the informal 'rules of the road' of cyberspace.

Debate on the use of profiles

To critically discuss the use of profiling (based on web-data mining) the production of profiles has to be made visible to the public. With openness as an important condition, there are three different options that need to be discussed: whether using certain data should be prohibited for profiling purposes, whether profiles should be disallowed for decision making or whether some decision-making methods should be rethought [Vedder, 2001]. For example, if we look at profiling done by medical insurance companies, we could question the mining of certain personal medical data, we could also question the use of those specific profiles, or we could rethink the use of enclosure and exclusion mechanisms used by medical insurance companies.

Legal measures

Laws reflect our ethical standards. Legal measures help to prevent chaos and offer a means of protection. It has been pointed out that current laws do not suffice. It is, however, disputed, whether or not legal measures are required as a solution to the problems concerning web mining. Some people expect data mining to challenge and possibly cross the boundaries of privacy legislation[38]. Others believe that existing laws offer sufficient protection, if only people would comply with the 'rules of play'[39]. Furthermore, most web data miners believe that possible dangers can be prevented by self-regulatory measures. However, [Scribbins, 2001] has found that some companies do breach their own privacy policies, she beliefs that effective enforcement and a right to redress for consumers are vital. [Clarke, 1998] also states that pure self-regulation has been demonstrated time and time again not to work. He believes there is an urgent need for self-regulatory codes to be given legislative stiffening. According to [Murray, 2001], legislation is essential, because self-regulation (and related solutions) needs a framework defined by legislation, for without such a framework there will be no basis for consumer confidence. He notes that it is rather difficult to legislate in detail, and therefore suggests that a new legislative strategy is needed dealing in frameworks and principles, and including a process for applying these principles to concrete cases.

Yet a problem arises, when considering how to monitor possible illegal or unethical web mining activities. The mere fact that the Internet is global creates numerous difficulties with regards to law enforcement. Many laws could prove meaningless and unenforceable. Furthermore, many data marketeers equate their ethical obligations with following the letter of the law [Wilder, 2001]. It is precisely this attitude of companies primarily describing ethical business practices in terms of complying with existing laws, that makes legal measures a 'dangerous' solution. People should never solely focus on the written rules. Moral awareness is at least as important as legal awareness, if not more important.

38 For instance [Borking, 1998].

39 [Artz, 2000] states that privacy of web users should be ensured by the Internet service provider (including site owners). According to him a stricter complyiance with current legislation is all it takes. In order to do this they have listed some rules of play.

Table 3
Possible solutions for the different types of web mining.

content and structure mining	usage mining
collective level	*collective level*
create awareness	privacy empowering web mining tools
monitor web content and structure mining activities	publish privacy policy
debate on use of profiles	web quality seal
	create awareness
suggestion:	monitor web usage mining activities
create 'disallow-mining standard'	debate on use of profiles
individual level	*individual level*
be cautious	be cautious
	use PET's
use 'disallow-mining standard'	check privacy policies and seals

When applying these solutions to the specific problems of the two web mining categories, we can classify the solutions as illustrated by Table 3. A mixture of technical and non-technical solutions at both the individual and the collective level is probably required even to begin solving some of the problems presented. But to what extend can the problems really be solved in both web mining categories?

Content and structure mining

Looking at the possible misuse of web content and structure data and the described solutions, we have to conclude that there is little that can be done to limit the dangers. Legal measures could provide a baseline level for better ways to handle the problems. However, because the mining of personal data published on the web is not presently prevented by legal measures, monitoring web content mining activities seems to be a useful additional solution. Notice, that this is difficult to actually implement. The monitoring should be done by a non-partial organization and it will only work if businesses co-operate and agree to give insight in their web mining activities. Without the miners co-operation it would be quite difficult to monitor web mining activities, because of the fact that they can be concealed so easily. Still, even with a well working monitoring 'system', there is no actual prevention of web mining damage. Therefore a discussion on the social and economical circumstances that make group profiling attractive could be fruitful. The discussion on rethinking the use of profiling is however a difficult one. Aside from the practical difficulties to make sure that certain data or profiles will no longer be used, determining what data or profiles are not to be used is also rather complicated. There are so many possible situations that would have to be considered, not to mention the unpredictable situations and patterns that can arise when applying web mining. Still the

essence of this debate on the use of profiles is quite important. Not only should people continue to be critical about using profiles, people should also discuss whether or not certain enclosure and exclusion mechanisms are just. Next to the advice for web users to be cautious when publishing personal information on the web, there is the collective 'obligation' to make people more aware of the dangers. This is not just about creating awareness amongst web users, but also about appealing to the moral sensitivity of web data miners.

This discussion on the practical value of these solutions also holds for mining web usage data. The interaction of a user with the web can, however, also be protected by some of the other solutions.

Usage mining

There are several tools that can help to protect a web user's privacy, while surfing on the web. A privacy empowering tool could enable the web user to make informed choices. It can help users understand privacy policies and it provides the option of informed consent. However, such a tool alone does not support all basic provisions of the EU data protection directive. There are quite a few skeptics who believe that many of these so-called privacy empowering tools are designed rather to facilitate data sharing than to protect users [Scribbins, 2001]. Consequently these tools do not offer a complete and satisfying solution for everybody. In addition, seal programs and third party monitoring help to ensure that sites comply with their policies. There are some people who remain skeptical about quality seal programs, because of the commercial interests. According to them the web quality seal programs have failed to require sufficiently high standards or to censure companies when they violate privacy. Ultimately it is up to individuals whether or not they trust the seals and privacy policies.
Individuals can to some extend protect themselves against web usage mining misuse by being cautious, when an on-line organization asks for personal data, and by using privacy enhancing technologies. Anonymity tools reduce the quantity of information revealed while browsing, just like cookies and ad rejecting tools. Notice, however, that if all web users were to ban ads, a lot of on-line services would be lost. Many search engines, web portals, news sites, and so on obtain their profits from selling on-line advertising spaces. Clear privacy policies, seal programs and privacy empowering tools can help to win a user's trust and perhaps make a user decide not to reject the ads and cookies from a web site.

For all of these solutions openness on web mining activities is required. Consumers have to be informed that data mining is being used by certain businesses and that data about them is currently being mined in ways that they probably had not explicitly authorized. Only then are they able to make

informed choices. Furthermore, as legal measures and ethical debates usually take some time before being effective, technical solutions seem to provide web users with direct possibilities to actively protect themselves against possible misuse. A web user can use different kinds of PET's to prevent on-line tracking. However, technical measures concerning the protection of content and structure mining misuse are not yet at hand. In this article a suggestion is made on a possible solution: creating a 'disallow-mining' standard.

Openness on the mining of content and structure data would imply that an individual should be able to determine whether or not (for example) the data on his homepage is allowed to be mined and if so, whether he should be informed about it.[40] A possible way to do this is by creating an automatic 'disallow-mining' standard. As asking each individual for his consent would be such a time-consuming matter, it will not be done by content miners. There might, however, be a way to automatically check for consent, based on techniques used by search engines. Search engines use web agents, also known as robots, to create the indexes for their search databases, the so-called 'spidering'. Robots.txt is a file that web agents look in for information on how the site is to be catalogued. It is a text file that defines what documents and / or directories are forbidden to index. Perhaps analogous to this, something like a 'mining.text' file could be created, which a content mining tool would check before mining the content of the site. In this file a site owner could state whether or not the personal information may be mined for certain purposes. As the robots.txt file already provides users with the option not to be indexed by search engines, it might also be possible to add an extra section, which states whether the site may be mined for other purposes. This solution only works for people who have access to the document root of their web site, because the robots.txt file is placed in the document root of the server. However, there are a lot of people who publish their homepage on a web space where they do not have root access. In those cases the site manager could for instance choose to alter the robots.txt file in case a user would have specific wishes. There is, however, also an option to direct the web agents on a per-page basis. Every HTML document contains a heading section in which meta data on the document (such as keywords, a description of the content and so on) can be included, so-called meta-tags. Within the meta-tags of each HTML document one can specify whether or not a robot is allowed to index the page and submit it into a search engine. So perhaps it would be possible to add a 'disallow-mining' option to the HTML standard. That way an individual could specify his wishes in the meta-tags of his own homepage.

Of course such techniques would only work, if it became a standard (such as robots.txt and meta-tags) and widely accepted by web content and structure miners. Notice, however, that some robots will simply ignore the meta-tags,

40 Not taking into account the fact that a person is not always aware of all the data that is published about him on the Web.

because of the fact that those tags are often misused by page owners, who want to get a higher ranking in a search index. Robots may also ignore the robots.txt file, or purposely load the documents that the file marks as disallowed. Therefore, robot exclusion files and meta-tags should not be regarded as a security measure. Misuse would still be possible, but organizations that have no intention of doing any harm can respect the person's wishes, without having to consult every individual first.

None of these individual and technical solutions address the protection of privacy and individuality once information is collected, and the problem of the nature of group profiles (the fact that one individual has little or no influence on the entire profile). As mentioned previously, even if an individual makes sure his data will not be used in data mining analyses, profiles derived from those analyses could still be projected on him. Therefore, individuals have limited possibilities to protect themselves. They could choose to combine forces and make sure that personal data is systematically refused by large groups of people. A large scaled refusal of data would, however, also block all the possible advantages of group profiling. Clearly, in order to benefit from web mining as much as possible, while causing no, or as little as possible harm, a combined solution-package is needed [41]. Or as [Clarke, 1998, page 48] states: "effective protection is dependent on a multi-partite, tiered framework, in which layers of technology, organizational practices and law combine to ensure reasonable behavior". To keep up with the ever-emerging technical changes, ongoing debates on the ethical issues are an essential part of this combined solution-package and can help prevent possible future damage.

4.1.7 Closing remarks

As the World Wide Web becomes an increasingly important part of modern society, organizations of all kinds are engaged in efforts to use web mining technology for varied purposes. Many of those purposes are of a commercial nature, for there is money to be made in the collection and intelligent analyses of information about people. Generally, web miners benefit most from web mining, while web users are facing the dangers.

The number of people that are touched and possibly affected by this technology is huge and growing rapidly. In the future the web use is likely to increase and so will the amount of valuable web data. Recent research shows there were around 1,5 billion Web pages in the beginning of the year 2000, an 88% increase from 1998. This suggests that 1,9 million Web pages are created each day. The number is expected to hit 8 billion in 2002, exceeding the world's population [Lake, 2000]. Although the impact of web mining should be of every web user's concern, there

41 [Johnson, 2001, pages 132 and 135] refers to this as a 'many-pronged approach'.

is no reason for people to panic. After having spoken to some web mining experts[42], it became clear that this technique is not yet being used to its full potential and that there is no clear indication of web data being misused to such an extent that people are actually hurt by it. Indeed, I mentioned that one of the dangers lies in the hidden way in which web mining can be used. Companies can cover up their ultimate goals quite easily, when they obtain certain bits of information. And one might get somewhat worried, when reading some of the web sites that offer web mining tools. Slogans like: 'The Internet is your database', 'Increasing the value of EVERY customer interaction', 'Turning data chaos into profit', and sentences like: "...to foster long-term customer relationships you need more than the analysis of log files and click stream behavior. We believe e-retailers and content providers need to know what their customers' values are and how and where they live", or "There is real data on the Internet, including addresses, phone numbers, calendars, prices, job listings", are found on web sites of companies that offer web mining services. Although this does indicate that the privacy and individuality of web users is threatened, there is still no reason for panic. As web mining is in an early stage of development, there are things that can be done to guide this technique in a socially acceptable direction. The solutions discussed in the former section can contribute to the responsible and well-considered development and application of web mining. As ethical issues will grow as rapidly as the technology, ethical considerations should be an integrated and essential part of this development process instead of something peripheral. It is probably impossible to develop comprehensive ethical guidelines covering every possible misuse. This is all the more reason to realize the seriousness of the dangers and to continuously discuss these ethical issues. This is a joint responsibility for web miners (both adopters and developers), web users and governments.

Acknowledgements

Many thanks go to the web mining experts participating in my (limited) field study including: Annius Groenink (Eidetica), Nils Rooijmans (ilse Media B.V.), Marten Den Uyl (Sentient Machine Research B.V.), Bert Verelst (Cyveillance International), Martijn Wiertz and Tom Khabaza (SPSS Inc.), Rob de Wit (WiseGuys Internet B.V.) and Bas Zinsmeister (Bolesian B.V.). Not only did they help me to gain insight into the existing and possible applications of web mining, they have also pointed out some critical arguments against the seriousness of the ethical issues concerning web mining.

Finally, I would like to thank Lambèr Royakkers, Anton Vedder and Jeroen Meij for their critical reviews and for guiding me during my Master's thesis research period.

42 It has to be noted that most of this (limited) field research is based on interviews with people from Dutch web mining companies (with the exception of SPSS and Cyveillance), so no worldwide conclusions can be drawn. Still the Dutch situation might give a first impression on how the web mining field is emerging. More importantly, the ethical issues are of global value (in the western world).

References

- Artz, M.J.T., M.M.M. van Eijk. (2000). Klant in het web. Privacywaarborgen voor internettoegang. Achtergrondstudies en Verkenningen 17. Registratiekamer, Den Haag
- Borking, J., M. Artz, L. van Almelo. (1998). Gouden bergen van gegevens. Over datawarehousing, datamining en privacy. Achtergrondstudies en Verkenningen 10. Registratiekamer, Den Haag
- Büchner, A.G., S.S. Anand, M.D. Mulvenna, J.G. Hughes. (1999). Discovering Internet Marketing Intelligence through Web Log Mining. Proceedings Unicom99 Data Mining & Datawarehousing: Realizing the full Value of Business Data. pp127-138
- Clarke, R. (1994). 'Profiling' and its privacy implications. Privacy Law & Policy Reporter **1**:128
- Clarke, R. (1998). Platform for privacy preferences: a critique. Privacy Law & Policy Reporter **5** (3):46-48. http://www.anu.edu.au/people/Roger.Clarke/DV/P3PCrit.html
- Custers, B. (2001). Data mining and group profiling on the Internet. In: A. Vedder (ed.). Ethics and the Internet. Intersentia, Antwerpen. pp87-104
- Fayyad, U.M., G. Piatetsky-Shapiro, P. Smyth, (1996a). From Data Mining to Knowledge Discovery: An Overview. Advances in Knowledge Discovery and Data Mining. AAAI Press/The MIT Press, Cambridge, Massachusetts
- Fayyad, U.M., G. Piatetsky-Shapiro, P. Smyth. (1996b). The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of the ACM **39** (11):27-34
- Fulda, J.S. (1999). Solution to a Philosophical Problem concerning Data Mining. Computers and Society **29** (4):6-7
- Holsheimer, M. (1999). Datamining ontdekt waardevolle informatie in databases. Privacy & Informatie **3**:100-104
- Houben, G-J. (2002). Mining for Adaptive Hypermedia. See other part of this book
- Johnson, D.G. (2001). Computer ethics. Prentice-Hall, New Jersey
- Khabaza, T. (2000). As E-asy as falling off a web-log: data mining hits the web. Proceedings of the Fourth International Conference on the Practical Applications of Knowledge Discovery and Data Mining, Manchester UK. The Practical Application Company. http://www.mining.dk/SPSS/Nyheder/as_easy.htm
- Kosala R., H. Blockeel. (2000). Web Mining Research: A Survey. ACM SIGKDD **2** (1):1-15
- Kosala, R., H. Blockeel, F. Neven. (2002). See other part of this book
- Lake, D. (2000). The web: Growing by 2 Million Pages a Day. The Standard (an on-line magazine). 28 Feb. http://www.thestandard.com/article/0,1902,12329,00.html

- Madria, S.K., S.S. Bhowmick, W.-K. Ng, E.P. Lim. (2002). Research Issues in Web mining. Lecture Notes in Computer Science **1676**:303-312
- Mobasher, B., R. Cooley, J. Srivastava. (2000). Automatic personalization based on web usage mining. Communications of the ACM **43** (8):142-151
- Mobasher, B., H. Dai, T. Luo, Y. Sun, J. Zhu, (2000). Integrating web usage and content mining for more effective personalization. Proceedings of the International Conference on E-Commerce and Web Technologies (ECWeb2000). http://maya.cs.depaul.edu/~mobasher/papers/ecweb2000.pdf
- Mulvenna, M.D., S.S. Anand, A.G. Büchner. (2000). Personalization on the net using web mining. Communications of the ACM **43** (8):123-125
- Murray, J. (2001). Electronic commerce – Regulation with a light touch. Challenge Europe (On-line journal from the European Policy Centre (EPC)). 28 March. http://www.theepc.be/Challenge_Europe/memo.asp?ID=424
- Nissenbaum, H. (1997). Toward an approach to privacy in public: challenges of information technology. Ethics & Behavior **7** (3):207-220
- Putten, P. van der. (1999). Graven naar Klantgegevens. Informatie en Informatiebeleid **17** (2)
- Schreuders, E. (2001). Data mining, de toetsing van beslisregels & privacy. Een juridische Odyssee naar een procedure om het toepassen van beslisregels te kunnen toetsen. Doctoral Thesis. University of Brabant, Tilburg, The Netherlands. http://www.weblex.nl/eric/bestanden/index2.htm
- Scribbins, K. (2001). Privacy@net, An international comparative study of consumer privacy on the internet. Consumers International, London. http://www.consumersinternational.org/news/pressreleases/fprivreport.pdf
- Spiliopoulou, M. (2000). Web usage mining for web site evaluation: making a site better fit its users. Communications of the ACM **43** (8):127-134
- Srivastava, J., R. Cooley, M. Deshpande, P.N. Tan. (2000). Web usage mining: Discovery and applications of usage patterns from web data. ACM SIGKDD **1** (2):12-23
- Tavani, H.T. (1999a). Informational privacy, data mining, and the internet. Ethics and Information Technology **1**:137-145
- Tavani, H.T. (1999b). KDD, data mining, and the challenge for normative privacy. Ethics and Information Technology **1**:265-273
- Vedder, A. (1998). Het einde van de individualiteit? Datamining groepsprofilering en de vermeerdering van brute pech en dom geluk, Privacy & Informatie **3**:115-120
- Vedder, A. (1999). KDD: The challenge to individualism. Ethics and Information Technology **1**:275-281

- Vedder, A. (2000). Privacy and Confidentiality. Medical data, new information technologies, and the need for normative principles other than privacy rules. Law and Medicine **3**:441-459
- Vedder, A., R. Holtmaat (ed.) (2001). Discriminatiegronden in het informatie-tijdperk, De toekomst van gelijkheid: de juridische en maatschappelijke inbedding van de gelijkbehandelingsnorm. Kluwer, Deventer, The Netherlands
- Vos, H. (1995). Filosofie van de moraal. Inzicht in moraal en ethiek. Aula Spectrum, Utrecht, The Netherlands
- Weckert, J. (2000). Computer ethics: future directions. Presented at the ARC Special Research Centre for Applied Philosophy and Public Ethics. Charles Sturt University in Canberra and The University of Melbourne. http://www.acs.org.au/act/events/2000acs4.html
- Wilder, C., J. Soat. (2001). The ethics of data. Informationweek.com (online magazine), May 14. http://www.informationweek.com/837/dataethics.htm
- Xiaohe, L. (1998). On economic and ethical value. The online journal of ethics **2** (1). http://www.depaul.edu/ethics/evalue.html

4

4.2 Legal issues of data mining

As we have seen in the previous chapter, data mining may involve ethical issues. In a legal perspective, these issues may concern (mainly informational) privacy, anti-discrimination and criminal law enforcement regulations. Or, to take another approach, we are concerned with data collection, legitimacy of derived rules and legitimacy of information within data mining operations.
In the next sections we will focus on general principles, valid in most democratic societies regardless of the juridical system of the particular region. These offer a starting point, but no text in this chapter is to be interpreted as legal advice. Concerning legal issues, the applicable laws and regulations should be referred to.
On the CD-rom, several legal directives are present, including a workbook [PDT, 2001] to assess actual the specific business situation in view of the OECD fair information practices discussed below. The OECD guidelines are also included on the CD-rom.

4.2.1 Fair information practices

Ann Cavoukian[1]

Around the world, virtually all privacy legislation, and the policies, guidelines, or codes of conduct used by non-government organizations, have been derived from the set of principles established in 1980 by the Organization for Economic Co-operation and Development (OECD). These principles are often referred to as 'fair information practices' and cover eight specific areas of data protection (or informational privacy):

1 Collection limitation.
2 Data quality.
3 Purpose specification.
4 Use limitation.
5 Security safeguards.
6 Openness.
7 Individual participation.
8 Accountability.

Essentially, these eight principles of data protection or fair information practices codify how personal data should be protected. At the core of these principles is the concept of personal control — the ability of an individual to maintain some degree of control over the use and dissemination of his or her personal information.

Concerns about informational privacy generally relate to the manner in which personal information is collected, used and disclosed. When a business collects information without the knowledge or consent of the individual to whom the information relates, or uses that information in ways that are not known to the individual, or discloses the information without the consent of the individual, informational privacy may be violated.

Data mining is a growing business activity, but from the perspective of fair information practices, is privacy in jeopardy? To determine this, we reviewed data mining from a fair information practices perspective. As discussed below, we have identified issues with five of these principles.

Data Quality Principle

Personal data should be relevant to the purposes for which they are to be used, and, to the extent necessary for those purposes, should be accurate, complete, and up-to-date.

1 A. Cavoukian, Ph D, Information and Privacy Commissioner, Ontario, Canada, http://www.ipc.on.ca/

Any form of data analysis is only as good as the data itself. Data mining operations involve the use of massive amounts of data from a variety of sources: these data could have originated from old, current, accurate or inaccurate, internal or external sources. Not only should the data be accurate, but the accuracy of the data is also dependent on the input accuracy (data entry), and the steps taken (if in fact taken), to ensure that the data being analyzed are indeed 'clean'.

This requires a data mining operation to use a good data cleansing process to clean or scrub the data before mining explorations are executed. Otherwise, information will be inaccurate, incomplete or missing. If data are not properly cleansed, errors, inaccuracies and omissions will continue to intensify with subsequent applications. Above all else, consumers will not be in a position to request access to the data or make corrections, erasures or deletions, if, in the first instance, the data mining activities are not known to them.

Purpose Specification Principle

The purposes for which personal data are collected should be specified not later than at the time of data collection and the subsequent use limited to the fulfillment of those purposes or such others as are not incompatible with those purposes and as are specified on each occasion of change of purpose.

Use Limitation Principle

Personal data should not be disclosed, made available or otherwise used for purposes other than those specified in accordance with the Purpose Specification Principle except:

- with the consent of the data subject, or
- by the authority of law.

Purpose Specification means that the type of personal data an organization is permitted to collect is limited by the purpose of the collection. The basic rule is that data collected should be relevant and sufficient, but not excessive for the stated purpose. In other words, restraint should be exercised when personal data are collected. Use Limitation means that the purpose specified to the data subject (in this case, the consumer) at the time of the collection restricts the use of the information collected. Hence, the information collected may only be used for the specified purpose unless the data subject has provided consent for additional uses.

Data mining techniques allow information collected for one purpose to be used for other, secondary purposes. For example, if the primary purpose of the collection of transactional information is to permit a payment to be made for credit

card purposes, then using the information for other purposes, such as data mining, without having identified this purpose before or at the time of the collection, is in violation of both of the above principles. The primary purpose of the collection must be clearly understood by the consumer and identified at the time of the collection. Data mining, however, is a secondary, future use. As such, it requires the explicit consent of the data subject or consumer.

The Use Limitation Principle is perhaps the most difficult to address in the context of data mining or, indeed, a host of other applications that benefit from the subsequent use of data in ways never contemplated or anticipated at the time of the initial collection. Restricting the secondary uses of information will probably become the thorniest of the fair information practices to administer, for essentially one reason: At the time these principles were first developed (in the late 70s), the means by which to capitalize on the benefits and efficiencies of multiple uses of data were neither widely available nor inexpensive, thus facilitating the old 'silo' approach to the storage and segregated use of information.

With the advent of high speed computers, local area networks, powerful software techniques, massive information storage and analysis capabilities, neural networks, parallel processing, and the explosive use of the Internet, a new world is emerging. Change is now the norm, not the exception, and in the quickly evolving field of information technology, information practices must also keep pace, or run the risk of facing extinction. Take, for example, the new directions being taken intending to replace the information silos of old, with new concepts such as 'data integration' and 'data clustering'. If privacy advocates do not keep pace with these new developments, it will become increasingly difficult to advance options and solutions that can effectively balance privacy interests and new technology applications. Keeping pace will enable us to continue as players in this important arena, allowing us to engage in a meaningful dialogue on privacy and future information practices.

The challenge facing privacy advocates is to address these changes directly while preserving some semblance of meaningful data protection. For example, in the context of data mining, businesses could easily address this issue by adding the words 'data mining' as a primary purpose at the time of data collection — but would this truly constitute 'meaningful' data protection? Take another example: when applying for a new credit card, data mining could be added to the purposes for which the personal information collected on the application form would be used. But again, would this type of general, catch-all purpose be better than having no purpose at all? Possibly, but only marginally so.

The quandary we face with data mining is what suggestions to offer businesses

that could truly serve as a meaningful primary purpose. The reason for this lies in the very fact that, at its essence, a 'good' data mining program cannot, in advance, delineate what the primary purpose will be — its job is to sift through all the information available to unearth the unknown. Data mining is predicated on finding the unknown. The discovery model upon which it builds has no hypothesis — this is precisely what differentiates it from traditional forms of analysis. And with the falling cost of memory, the rising practice of data warehousing, and greatly enhanced processing speeds, the trend toward data mining will only increase.

The data miner does not know, cannot know, at the outset, what personal data will be of value or what relationships will emerge. Therefore, identifying a primary purpose at the beginning of the process, and then restricting one's use of the data to that purpose are the antithesis of a data mining exercise.

This presents a serious dilemma for privacy advocates, consumers, and businesses grappling with the privacy concerns embodied in an activity such as data mining. To summarize, the challenge lies in attempting to identify as a primary purpose, an as yet, unknown, secondary use. We offer some suggestions on how to address this issue in the next section.

Openness Principle

There should be a general policy of openness about developments, practices, and policies with respect to personal data. Means should be readily available of establishing the existence and nature of personal data, and the main purposes of their use, as well as the identity and usual residence of the data controller.

The principle of openness or transparency refers to the concept that people have the right to know what data about them have been collected, who has access to that data, and how the data are being used. Simply put, it means that people must be made aware of the conditions under which their information is being kept and used.

Data mining is not an open and transparent activity. It is invisible. Data mining technology makes it possible to analyze huge amounts of information about individuals — their buying habits, preferences, and whereabouts, at any point in time, without their knowledge or consent. Even consumers with a heightened sense of privacy about the use and circulation of their personal information would have no idea that the information they provided for the rental of a movie or a credit card transaction could be mined and a detailed profile of their preferences developed.

In order for the process to become open and transparent, consumers need to know that their personal information is being used in data mining activities. It is not reasonable to expect that the average consumer would be aware of data mining technologies. If consumers were made aware of data mining applications, then they could inquire about information assembled or compiled about them from the business with which they were transacting — 'information' meaning inferences, profiles and conclusions drawn or extracted from data mining practices.

Ultimately, openness and transparency engender an environment for consumers to act on their own behalf (should they so choose). Consumers could then make known to the businesses they were transacting with, their expectations about the collection, re-use, sale and resale of their personal information.

Individual Participation Principle

An individual should have the right:

a to obtain from a data controller, or otherwise, confirmation of whether or not the data controller has data relating to him;

b to have communicated to him, data relating to him i) within a reasonable time, ii) at a charge if any that is not excessive, iii) in a reasonable manner and iv) in a form that is readily intelligible to him;

c to be given reasons if a request made under subparagraph (a) and (b) is denied, and to be able to challenge such denial; and

d to challenge data relating to him and, if the challenge is successful, to have the data erased, rectified, completed or amended.

Data mining operations are extremely far removed from the point of transaction or the point of the collection of the personal information. As data mining is not openly apparent to the consumer, then the consumer is not aware of the existence of information gained through a data mining application. This prevents any opportunity to: 1) request access to the information, or 2) challenge the data and request that corrections, additions, or deletions be made.

Choices for consumers and businesses

Consumers

In the United States, media coverage of public concerns about informational privacy matters began around the start of this decade with the uproar that erupted over Lotus Marketplace: Households[2]. This was an early and perhaps defining demonstration of the public's sensitivity about informational privacy. In 1996, the Lexis-Nexis incident drew massive attention to how people feel about their personal information: Lexis-Nexis, an online information service in Dayton, Ohio

2 Lotus Marketplace: Households was a series of disks produced by Equifax and Lotus Development Corporation in 1990. On these disks (available to anyone for a price), were the names, addresses, buying habits and income information of roughly 120 million American consumers. Over 30,000 consumer enquiries and complaints lodged shortly after its release effectively cancelled the sale of the disks.

was accused of making social security numbers and other personal information widely available in its P-TRAK locator service. Then in 1997, after an electronic firestorm, America Online backed off of its plan to rent out its subscribers' telephone numbers [Smith, 1997]. In each of these cases, businesses quickly responded to a public outcry from their customers and either withdrew their products or changed their policies.

However, in order for consumers to react (and businesses to respond), consumers must have knowledge and awareness that something they could potentially choose to object to is actually occurring. The invisible nature of data mining (to the consumer) eliminates this possibility. In order for data mining to fall into line with fair information practices, the first step for consumers must be an awareness that any large business they are transacting with could be carrying out data mining activities. For some consumers, this knowledge will make no difference; for others, it will matter a great deal.

Once consumers are equipped with knowledge, it is up to each individual to decide for him or herself what matters, and based on that, what choices they want to make about assuming control over the uses of their personal information.

Concerned consumers can choose to take responsibility by informing businesses of their requirements and expectations regarding privacy. To assist in framing privacy-related questions relating to data mining, consumers may wish to consider the questions below. Then it is up to the consumer to decide what course of action, if any, to take.

Actions

As a consumer

- Do you expect to be informed of any additional purposes that your personal information may be used, beyond the primary purpose of the transaction?
- Do you expect the option to say 'no' to secondary or additional uses of your personal information, usually provided in the form of opting-out of permitting the use of your personal information for additional, secondary uses? Or, do you expect an opportunity to 'opt-in' to secondary uses?
- Do you expect a process to be in place that gives you the right to access any information a business has about you, at any point in time?
- Do you expect a process that permits you to challenge, and if successful, correct or amend any information held by a businesses about you, at any point in time?
- Do you expect an option to have your personal information anonymized for data mining purposes and/or, an option to conduct your transactions anonymously?

For those consumers who wish to have greater control over the use and circulation of their personal information, we suggest the following initiatives. Ask to see a business's privacy or confidentiality policy. Assess it against your expectations of how you want your personal information handled. If the policy does not meet your expectations, contact the business and inform it of your expectations. If no policy exists, inform the business that you expect respectful and fair handling of your personal information.
Give only the minimum amount of personal information needed to complete a transaction. If you are in doubt about the relevance of any information that is requested, ask questions about why it is needed, and ask that all of the uses of the requested information be identified.

Actions

Businesses

Businesses need a corporate will to adopt a culture of privacy — piece-meal or theoretical approaches will not be effective in responding to consumers' concerns. Ultimately, the impact of various technologies on privacy, including data mining, can only be averted by instilling a culture of privacy within the organization.

'Instilling a culture of privacy' means that businesses will have to tackle the conflict between the 'use limitation' principle and the secondary uses of personal information arising out of data mining. It may be advisable for businesses to provide a multiple choice opt-out selection whereby consumers are given three choices: the choice of not having their data mined at all; only having their data mined in-house; or having their data mined externally as well. (Studies have shown that less concern is expressed over the internal secondary uses of one's data by the company collecting the data, but far greater resistance to having data disclosed externally for use by unknown parties).

Is your business willing to:

- have a privacy strategy that is:
  - based on fair information practices and entrenched through tangible actions;
  - resourced throughout all facets of the organization;
  - evaluated and assessed so that ongoing adjustments and improvements can be made?
- have an open and transparent relationship with its customers?
- Do you inform your customers upfront as to how all information collected about them will be used and disclosed, and by whom?
- Do you have a process that makes it easy for customers to find out what personal information you have about them and a process to challenge any information that may be incorrect, incomplete, inaccurate or out-of-date?

- accept that some consumers do not want their personal information to be mined, and nuggets about their buying patterns extracted?
- Do you advise consumers of all uses of their personal information and give them a range of opt- out choices about data mining such as: 1) no data mining; 2) data mining internally; 3) data mining internally and externally. Or, for maximum choice and control, do you provide consumers with positive consent — an opportunity to 'opt-in' for specified secondary uses of their personal information?
- use privacy-enhancing technologies that can anonymize information and securely protect privacy?

A FINAL WORD

The need for protecting and managing personal information has been likened to the management of natural resources.

Personal information is a resource, exploited commercially but valued as an element of human dignity and enjoyment of one's private life. It is therefore to be protected and managed, not unlike the protection and management of other resources. As with early efforts to protect the environment in the absence of legislation, privacy protection currently relies on ancient common law principles that continue to adapt to new technological challenges to personal integrity, happiness and freedom. These principles have now found legislative expression in various statutes relating to environmental protection. Information, however, has some unique qualities in need of special regulatory and judicial attention [Lawson, 1992].

Looking ahead, consumers will not only want goods and services, but will increasingly want assurances that the information they provide to a business is, from a privacy perspective, protected. To deal with this need, a shared responsibility for the management of personal information will be essential, involving government, the business community and consumers. Only through shared responsibility, sustained by the business community through a culture of privacy, and strengthened by the voice of consumers, can personal information become a protected, managed and valued resource. We hope that this report will give all three parties — consumers, businesses, and government — incentives for action towards protecting personal information in the marketplace.

Finally, we believe that the tension between technology and privacy can be minimized if privacy safeguards are made a key consideration upfront, rather than as an afterthought. Although current data mining practices are somewhat beyond the 'upfront' stage, there is still time to ease this 'tension' before applications become widely commonplace. One short term approach, as suggested

earlier, may be for businesses to provide consumers with choices in the form of multiple selection opt- outs. To explore further solutions on how to address the 'primary purpose' dilemma that data mining presents, we are committed to an open exchange. We invite those of you with any ideas as to how to resolve this issue to contact us — we would welcome your comments and encourage an open dialogue.

References

- OECD. (1980). Guidelines on the Protection of Privacy and Transborder Flows of Personal Data
- Smith, R.E. (1997). Rapid-Response Time. Privacy Journal
- Clarke, R. (1997). Privacy and Dataveillance, and Organizational Strategy. http://www.anu.edu.au/people/Roger.Clarke/DV/PStrat.html. Viewed on October 9, 1997. This paper Presents a Framework to Guide Businesses and Governments towards Adopting a Strategic Approach to Privacy
- Lawson, I. (1992). Privacy and Free Enterprise: The Legal Protection of Personal Information In the Private Sector. (Ottawa: Public Interest Advocacy Centre. p442
- PDT. (2001). Privacy Diagnostic Tool (PDT). Version 1.0. Workbook. Privacy Commissioner/Ontario, Gardent Canada. PriceWaterhouseCoopers

4.2.2 Legitimacy of decision rules

Summary by Jeroen Meij

This is a summary of E. Schreuders' Ph D work on assessment of decision rules and privacy [Schreuders, 2001].

The decision rules generated from a data mining operation are (in a business environment) generally derived for business purposes.
Obviously, these decision rules and the actions that follow from their use, need to conform to legal boundaries. Essentially, this means the rule should be well founded, usable, allowed and acceptable. In the context of this Chapter we will try to provide some directions for their assessment in the light of privacy and anti discrimination legislation.

[Schreuders, 2001] formulates four elements that constitute a motivated decision rule:

1 The criteria of the decision rule.
2 The decision of the decision rule.
3 The motivation regarding the criteria.
4 The motivation regarding the decision.

For a juridical assessment, these eight questions can be combined with eight questions regarding the data mining processes and the decision rules originating from them.

1 What is the (business) purpose of the data processing?
2 What is the purpose of the decision rule, why is it needed and what will be the expected use?
3 Which data have been used, what are the selection criteria for accepting or rejecting certain data, what is the origin of the data, how have they been processed? (Involving cleaning, completing, errors, have errors been corrected or are they acceptable).
4 Which techniques have been used for analysis and why?
5 Have the rules been tested, how, and what was the outcome of the tests?
6 How is the decision rule formulated and on which actions, behavior or properties of person(s) is it based? What are possible uses and limitations of the rule and how do these fit into the purposes stated in 1 and 2.
7 Which are the business actions for which the decision rule is/will be used, and how does this relate to the purposes stated in 1 and 2 and the possible uses and limitations 6.
8 If the rule is applied to an individual: in which way is it determined how the rule can be applied to an individual and how does this relate to the purposes in 1 and 2 and the possible uses and limitations of 6.

This method may help the judgment of a decision rule against actual applicable laws. Assessment of the composition and application of decision rules will concentrate on the question whether a justified or unjustified distinction is made. Informational privacy and privacy protection are main points of attention. Schreuders emphasizes that, especially for privacy issues, it will not be possible to cover all cases in legal rules beforehand. New cases might create additional legislative demands, requiring political decisions periodically.

Reference

– Schreuders, E. (2001). Data mining, de toetsing van beslisregels & privacy. Een juridische Odyssee naar een procedure om het toepassen van beslisregels te kunnen toetsen. Ph D Thesis. Tilburg University, The Netherlands

4.2.3 Regulation of criminal law enforcement

Jan Grijpink[3], *Joop Verbeek*[4], *Jeroen Meij*

In this section, we consider aspects of privacy and anti-discrimination in the context of special regulations for criminal law enforcement. Its purpose is to provide a framework for discussion, the text should not be interpreted as legal advice.

Information used by criminal law enforcement

In police investigations, available databases might be used, but new data is often also collected and stored in databases. In this context, it is useful to distinguish between five types of databases:

1 Police databases (official registers) aimed at criminal law enforcement (such as criminal records, observation records, wiretap transcripts). This type can be subdivided into permanent and temporary databases.
2 Police databases not explicitly aimed at criminal law enforcement (such as license databases for liquor and firearms).
3 External databases of public authorities or private organizations with a public task (i.e. municipal registers, tax registers, chamber of commerce register).
4 Open sources (i.e. libraries, newspapers, the Internet).
5 External databases of private organizations (private and company databases).

A wide variety of sources may be used to gather data, delivering valuable investigative information.

An important observation is that the official use of information by criminal law enforcement is allowed only if:

– *The information is acquired legitimately.* Generally speaking, permission of a higher authority (for instance the prosecutor's office) is required for the acquisition of data, even if the data is handed over with consent of the owner of the database.
– *The information is processed legitimately.* This encompasses any analytical steps of the data mining activity, which are discussed in following subsections.
– *The information is stored legitimately*. Criminal law enforcement authorities are not necessarily permitted to store all forms of data. For instance, personal data from open sources may be observed, but not stored for law enforcement activities unless the person is suspected of having committed a criminal offence.

3 Dr Mr J.H.A.M. Grijpink, Principal advisor, Directorate of Strategy Development, Dutch Ministry of Justice, The Hague, The Netherlands

4 Mr P.G.J.M. Verbeek, Verbeek@cs.unimaas.nl, Institute for Knowledge and Agent Technology (IKAT), Department of Computer Science, Faculty of General Sciences, Universiteit Maastricht, Maastricht, The Netherlands

Privacy protection and exceptions

The OECD guidelines outlined in 4.2.1 provide a general framework for privacy protection. For the implementation of the guidelines we have to look at the applicable international and national legislation. At the international level, there are three legal instruments we will discuss [See also Sietsma et al., 2002]:

- the International Covenant on Civil and Political Rights.[5]
- the European Convention for the Protection of Human Rights and fundamental Freedoms[6]
- the European Directive[7] of 15 June 1995 on the protection of personal data.

The international directives are in most cases implemented in national legislation. EU member state legislation conforms to European conventions and directives. [Privacy, 2001] gives a comprehensive overview of privacy protection in countries all over the world. We will briefly discuss the Dutch legislation in the last subsection.

The International Covenant on Civil and Political Rights

Article 17 of this Covenant states that "no one shall be subjected to arbitrary or unlawful interference with his privacy, family, home or correspondence". The second section of the article contains the condition that "everybody has the right to the protection of the law against such interference or attacks".

The European Convention for the Protection of Human Rights and Fundamental Freedoms

Article 8 of the European Convention guarantees the right to respect for private and family life, and is interesting because of its direct applicability. Violation of this right to privacy is only permitted when the violation is deemed necessary by law, in a number of listed cases such as prevention of crime or disorder, or protection of public safety.

The European Directive of 15 June 1995 on the protection of personal data

This directive demands from EU Member States to provide that personal data must be:

a processed fairly and lawfully;

b collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes. Further processing of data for historical, statistical or scientific purposes will not be considered as incompatible provided that Member States provide appropriate safeguards;

c adequate, relevant and not excessive in relation to the purposes for which they are collected and/or for which they are further processed;

d accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that data which are inaccurate or incomplete, having

5 UN Office of High Commissioner of Human Rights, International Covenant on Civil and Political Rights, entry into force on 23 March 1976.

6 Council of Europe, European Convention for the Protection of Human Rights and Fundamental Freedoms, entry into force 3 September 1953.

7 Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (included on the CD-rom). Official Journal L 281 , 23/11/1995.

regard to the purposes for which they were collected or for which they are further processed, are erased or rectified;

e kept in a form which permits identification of data subjects for no longer that is necessary for the purposes for which the data were collected or for which they are further processed. Member Sates are to lay down appropriate safeguards for personal data stored for longer periods for historical, statistical or scientific use.

The directive poses many restrictions and obligations related to personal data processing, for instance:

- It grants individuals in the data base right of access to their data, and insight into the process.
- Some exceptions noted, the subject in the data base must have given explicit consent for his data to be processed.
- The directive prohibits the processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, and the processing of data concerning health or sex life.

This directive is applicable whenever processing of personal data takes place, which is often the case in data mining. However, according to a general juridical principle, general laws are superseded by special laws. Article 13 of the Directive enables states to make exceptions on the directive for the purpose of certain public interests such as state security and investigating and prosecuting crime. Data mining by criminal law enforcement it is covered by the exceptions of article 13 of the Directive. In addition, the Directive explicitly states that law enforcement and criminal law in general are not part of its scope of application[8].

Exceptions for criminal law enforcement

Three categories of data processing related with information discovery may be distinguished for criminal law enforcement [Sietsma et al., 2002]:

- verification.
- directed data mining.
- undirected data mining.

For each of these categories we can assess the restrictions posed by the covenant and directives discussed in the previous subsection. From this, some conclusions can be drawn, based on the general situation in the EU. Note that non EU nations may pose other restrictions.

8 See article 3 section 2 of the Directive.

Verification

Verification is performed within a criminal investigation to determine whether a hypothesis is true. According to the covenants and directives discussed in the previous subsection, data processing is generally allowed for verification purposes. Because there is a hypothesis concerning a known suspect, the use of type 1,2,3 and 4 databases is permitted. The use of type 5 databases without consent from the owner requires a judicial or court order. In most cases verification is not considered data mining according to the definition used in this book (see Section 6.1.2).

Directed data mining

Directed data mining is performed to obtain information on a specific crime, suspect or group of suspects. According to the covenants and directives of the previous subsection, data mining is usually allowed on type 1 databases, and under special conditions[9] in the context of a criminal investigation on all five database types. In directed data mining, the use of type 5 databases requires that these databases are voluntarily handed over to the criminal law enforcement authorities.

Undirected data mining

Undirected data mining is performed to improve the general information position, to gain more general insight, for instance, looking for patterns related to generally described crime types or societal crime related patterns. In principle, only the use of type 1 (law enforcement databases) is allowed for this type of data mining. The same purposes may be served by outsourcing the data mining to scientific research organizations, allowing the use of other types of databases (using anonimized data). For matters involving national security, the national security authorities could do the data mining within the boundaries of national security laws. Any conclusions useful for criminal justice can be handed over to these authorities.

Please note that the borders between the data mining categories used above may in practice be not so sharp, but more of a fuzzy nature. Again in this area legislation will develop stepwise, demanding regular efforts to adapt to changing technology and environments.

The situation for criminal law enforcement in the Netherlands

We will end this section with a discussion of the specific situation in The Netherlands. Details of database storing regulations are beyond the scope of this article, which is restricted to basic principles.
Article 10 of the Dutch constitution covers the protection of privacy. The article states that:

9 An example is art. 126 gg of the Code of Criminal Procedure in The Netherlands, applicable to undirected as well as directed data mining, see the next subsection.

– Everyone has the right to respect for his privacy, without prejudice to restrictions laid down by or pursuant to an act of parliament.
– Rules to protect privacy shall be laid down by act of parliament in connection with the recording and dissemination of personal data.
– Rules concerning the rights of persons to be informed of data recorded concerning them and of the use that is made thereof, and to have such data corrected shall be laid down by act of parliament.

On the basis of this article, the registration and processing of personal data has to be provided for by act of parliament, currently implemented in the Dutch data protection law, the WBP[10]. Privacy guidelines like those from the OECD fair information practices have been implemented in this law. Specific regulations for criminal justice and law enforcement have been formulated in the Code of criminal procedure[11], the Police registers act, WPolr[12] and several other laws.

The legislative authorities intended to arrange all the issues concerning police databases in the WPolr and its related legislation. Therefore, related articles in other laws have to correspond with the WPolr. There are many reasons for the creation of a separate law to arrange the processing of personal data by the police. We mention three:
– When police authorities gather personal information, the subject generally does not cooperate or does not know about the fact that he is being investigated.
– The protection of privacy is especially of importance when one looks at the sensitive place that is taken by the police in the relation between the state and its citizens.
– A separate law offers ample opportunity to log any sharing or editing of data, which guarantees better privacy protection.

As stated in the previous subsection, data mining not directed at a specific suspect or crime is restricted to criminal law enforcement databases. However in the context of the preparations for a criminal investigation, art. 126 gg of the Code of criminal procedure is applicable.

This article gives the prosecutor the right to authorize the police to investigate external data collections handed over to the authorities voluntarily to enable them to prepare a criminal investigation. This handing over of personal data is generally not allowed according to the WBP, because this does not conform to the purpose of the original data collection (this being the basic criterion of the WBP). For this special purpose the WBP contains a clause (art. 43) which states that any organisation is authorized to give their databases voluntarily to the criminal law enforcement authorities for the purpose of criminal investigations.

10 Wet Bescherming Persoonsgegevens.

11 Wetboek van strafvordering.

12 Wet politieregisters.

References

- Privacy (2002), Privacy International web site. http://www.privacyinternational.org
- Schreuders, E. (2001). Data mining, de toetsing van beslisregels & privacy. Een juridische Odyssee naar een procedure om het toepassen van beslisregels te kunnen toetsen. Ph D Thesis. Tilburg University, The Netherlands
- Sietsma, R., J.P.G.M. Verbeek, H.J. van den Herik (2002). Datamining en opsporing. IT*e*R series, NWO, The Hague

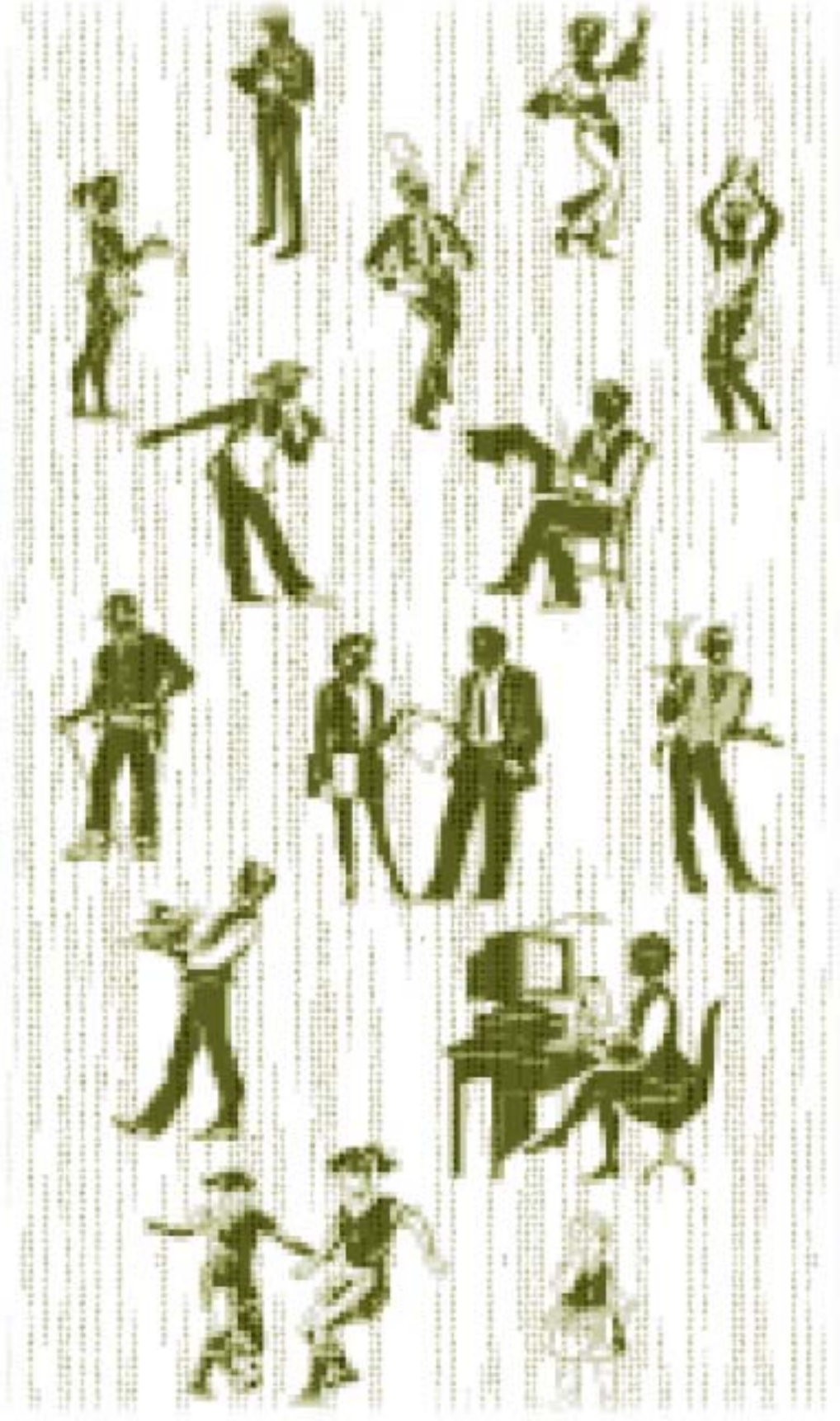
5 The Perspective of the Individual

5.1 Introduction

Jeroen Meij

This part of the book is centered around the individual use of data mining related techniques. There is no differentiation between an individual in a professional situation and an individual in a private situation. Both will need to find specific information, whether it is enclosed in text, hypertext, video, images or audio. This view is supported by the notion that the boundaries of individual and professional information acquisition are becoming more and more blurred. This is a logical consequence of the rise of home offices and of office homes. As many knowledge workers tend to be highly interested in the subjects that they are involved in, they are also extending their activities to their private lives. Similarly, people spending much time in the office will sometimes use their breaks to acquire information of personal interest.
Not surprisingly, this part focuses on content analysis, with connections to data conservation, learning and adaptive systems.

There is also another issue that is worth our attention: what is all this data about? The next chapter will briefly go into the kinds of data that are of interest to the individual user. We will try to give an impression not only of the data that will be available about the individual, but also of the data that is available to the individual.



5

5.2
Data acquisition and registration

The individual will shortly be confronted with an increase in data about him/her self in practically all areas of life. To create some insight into what this will eventually mean, we will start with an overview of the data that is being collected about individuals. On the other hand, individuals also have the benefit of access to ever growing quantities of data to suit their own interests. The second part of this chapter will describe the various possible sources.

5.2.1 Data about the individual

In modern society, registration is omnipresent. From womb to tomb, thousands of databases are updated with data from our actions in life. We would like to name a few, starting with the present and adding some expectations of the future. A survey in the Netherlands found that on average every citizen is present in 800 to 1,000 databases.

Some examples

Transaction data: shopping, money transfers, etc.

Whenever you pay with your credit card or bank card, you identify yourself, and your identity can be linked to your purchase. This can also be true of member cards, bonus cards, etc. Needless to say, banks also keep a list of their transactions. The customer gets a paper copy of these data: a sales slip or bank receipt.

Telecommunication: phone calls, GSM data

An extensive log is made by the phone companies, which includes time, call duration and numbers called. The same applies for calls from mobile phones, where the location of the mobile phone can also be included. Usually the customer has access to some of the transaction data through specified bills. Recently on-line information systems provide up to date transaction data for the customer.

Personal data by registration forms and web forms

Many hardware and software vendors ask people to supply personal data for registration of their purchase. Some web sites require a membership before access is allowed, for which the visitor has to provide some or (on occasion) very detailed information about himself, preferences, etc, including income and education. The use of this data is limited by legislation depending on the country and also by the privacy statement on the web site. After submission, the visitor usually has no access to or control over the supplied information.

Web clicks

A web site owner can log the path a visitor followed on his web site. Through cookies he can identify a user, when he turns back to the site.

Some future examples

Examples that may be seen in the future include an individual's personal genetic sequence, the recorded output of surveillance cameras, traffic monitoring data, etc.

5.2.2 Data for the individual

Typically, a television broadcast is a data stream that an individual could convert to knowledge. It contains audio and video, sometimes combined with text. Currently, a television signal is analog, and is not available at user requested times. It can be stored in analog form to overcome this problem. The data cannot be accessed randomly, is not indexed or annotated and not searchable. Also, the time base of the information is fixed. At the moment, broadcasts and music are mainly collected and stored for recreational purposes.
Most people have a wide collection of information on paper, such as magazines, newspaper clippings, letters, user manuals that are set aside, because it is expected that they will be of use at a later time. Since the rise of the Internet, this collection is expanded by a list of URL's of interesting web sites. Important information is usually printed or stored in cupboards, boxes, or on local hard disks.

As many data and information types enter the digital realm, it becomes possible to manage and integrate the knowledge and information stored in various data formats. Getting digital is one trend, storing your data on-line is another. Many sites are now offering storage space for music, photographs, or other files. Whether this trend will continue is a question of trust, access speed and cost.
As video enters the digital domain — digital video reorders are already commercially available — newsreels, web casts and documentaries can be stored and analyzed for later use. The same is possible for radio and web radio and other audio sources (recorded lectures, etc.).
Music can be analyzed to select the music that suits the users preference and mood at a certain time.
In the near future, we will experience a large increase in digitally available data sources. Some examples:

- digital libraries;
- digital reference sources (encyclopedias, etc.);
- personal preference data for the home environment ;
- personal biomedical data;
- personal food habits/preferences;
- personal entertainment habits/preferences;
- recorded conversations, such as audio or converted to text;
- contextual information and experiences, data picked up by advanced wearable sensor systems.

When reading the chapters of this part, these examples may help create an image of what will be the future uses of data mining from the perspective of the individual. We will pay attention to data conservation and maintenance in Chapter 5.3. Chapter 5.4 focuses on Text mining, followed by Multimedia mining in Chapter 5.5. Web mining is covered extensively in Chapter 5.6, followed by knowledge integration and learning aspects in Chapter 5.7. We end with some conclusions and expectations for the future.

5

5.3 Data Conservation and Maintenance

Trudi C. Noordermeer[1]

Introduction

The objectives of this article are to give an overview of digital archiving issues from the point of view of national libraries. It describes some important initiatives. Furthermore, it provides the state-of-affairs of the Depository System for Dutch Electronic Publications (DSEP or E-depot), which is developed by IBM and the National Library of the Netherlands, the Koninklijke Bibliotheek (KB).

Digital information and its storage media are vulnerable. Tapes, diskettes, CD-roms, hard drives and other storage devices break down over time, and eventually become brittle or less able to retain the digital information magnetically encoded on them. Archivists and librarians claim data stored on frequently used media such as tapes, floppy disks, CD's and other products may suddenly become unreadable over years, rather than decades or centuries.
With computer and software technology changing rapidly, the tools needed to read documents and data become obsolete quickly. If somebody has old tapes, or 8- or 5 $^1/_4$-inch floppy disks in their desk drawer, they may not be able to find the proper drives or software to read them.

1 Drs T.C. Noordermeer, Trudi.Noordermeer@kb.nl , Koninklijke Bibliotheek, National Library of The Netherlands, The Hague, The Netherlands

Operating systems, formats and software also get obsolete in a few years time. Sometimes software is backward compatible, but this is not always the case. Furthermore, the system hardware and storage media are all perishable and get obsolete or decay rather quickly. Therefore worldwide government organizations, national libraries and archives are investigating strategies and methods for digital preservation, including archiving of the Internet. Should people 'migrate' their data every few years to ensure that it is not lost or corrupted, or are there ways of emulating hardware and software to provide access to the original documents?
At this moment the issue of digital archiving cannot be solved in an adequate way, but worldwide people are becoming aware of the problem and many research projects are being carried out, and pilot systems built.

Definitions

This article concentrates on the strategy of a national library concerning the preservation of publications in electronic form which contains a substantial amount of text. Nowadays much information is available in digital form and the distinction between a text publication and an audiovisual document is sometimes rather hard to define. A multimedia publication contains text, stills, moving images and sound. The text document, which would be kept in the depository, could be originally published in electronic form ('born digital') or it might be an electronic publication containing digitized information. Furthermore, the document is made for broad distribution, which excludes e.g. personal web sites and email.
So, an electronic publication is a publication containing data meant for public use by electronic means, usually through the use of a combination of hardware, software and network facilities. An off-line publication is an electronic document which is bibliographically identifiable, stored in machine readable form on an electronic storage medium. Cd-rom, diskettes or floppy discs and magnetic tapes are examples. There are off-line monographs, like a CD-rom encyclopedia and off-line serials like a CD-rom journal. An on-line publication (or resource) is a bibliographically identifiable electronic document, stored in machine readable form on an electronic storage medium and available on-line. Examples are electronic journals, World Wide Web pages or on-line databases.
Nowadays, multimedia publications are produced containing a biography, a bibliography, photo's, animation, video and sound. It becomes increasingly difficult to distinguish between an audiovisual document and an electronic text publication. A movie with subtitling is considered to be audiovisual. A compact disc of a pop group with a video clip — consisting of moving images — is considered an audio CD. A CD-rom containing a biography, a bibliography, texts of the songs, sound, some video and photo's is considered to be a multimedia CD-rom publication. In short, an electronic publication must contain a 'considerable amount'

of text in order to be included in the national bibliography and taken into deposit of a national library. Of course, some national libraries also take audio-visual publications into deposit.

Problem area

The judicial, organizational, financial and technological questions involved in organizing a DSEP or E-depository are complicated. These include the criteria and method (push or pull) for the selection of electronic off-line and on-line publications, the workflow in the national library, the fields required for bibliographical and technical description, meta-data and unique identification, the variety of types of electronic publications and the fact that hardware (computers) and software (operating systems, application standards) quickly become obsolete. Carriers like magnetic tapes and diskettes have a short life expectancy. For CD-rom the life expectancy is somewhat better, but still, in view of the objectives of the national libraries, not long enough. The Internet and the World Wide Web are extremely volatile. So what are the best methods for web archiving?

Another major challenge national libraries are facing is how documents originally published in electronic form can be preserved and kept accessible for a very long time. The commercial life cycle of an electronic publication is usually short, because publishers obviously no longer have any (commercial), interest when the software is out of date or when the hardware for which a publication was made is obsolete.
National libraries worldwide have recognized that electronic publications should also be preserved for the remote future, just like printed publications and the national libraries should become digital archives. This means that the national library:

- defines selection criteria;
- receives the electronic publication (off-line CD-roms or diskettes by regular mail or on-line with FTP, email or web harvesting);
- will make bibliographical and technical meta-data;
- will publish this description in the national bibliography;
- will handle meta-data and unique identifiers;
- will have to migrate (with the publisher's permission) the contents to another carrier or format, when the hardware and software get obsolete or methods for emulation are developed;
- will guarantee authenticity and integrity, which means that a certain specified document is the 'real' version related to the meta-data;
- will give access (on site) to the end users now and in the remote future.

Digital preservation strategies

While digital technologies are enabling information to be created, manipulated, disseminated, located and stored with increasing ease, preserving access to this information poses a significant challenge. Unless preservation strategies are actively employed, this information will rapidly become inaccessible. The choice of strategy will depend upon the nature of the material and what aspects are to be retained.

Refreshing (copying information without changing it) offers a short-term solution for preserving access to digital material by ensuring that information is stored on newer media, before the old media deteriorates beyond the point at which the information can be retrieved.

The migration of digital information from one hardware/software configuration to another, or from one generation of computer technology to a later one, offers one method of dealing with technological obsolescence.

While adherence to standards will assist in preserving access to digital information, it must be recognized that technological standards themselves are evolving rapidly.

Potentially, technology emulation offers substantial benefits in preserving the functionality and integrity of digital objects. However, its practical benefits for this application have not yet been well demonstrated.

Encapsulation, a technique of grouping digital objects together with anything else necessary to provide access to the object, has been proposed by a number of researchers as a useful strategy in conjunction with other digital preservation methods.

The importance of documentation as a tool to assist in preserving digital material is universally agreed upon. In addition to the meta-data necessary for resource discovery, other sorts of meta-data, including preservation meta-data, describing the software, hardware and management requirements of the digital material, will provide essential information for preservation.

The requirement to keep every version of all software and hardware, operating systems and manuals, as well as relevant skills generally, makes the preservation of obsolete technologies an infeasible strategy.

Various frameworks designed to assist in managing the preservation of digital material have been developed. These include tools designed for assisting in the development of digital preservation strategies. Often these will entail the identification of the various stages at which the provision of long-term access should be considered.

Tasks for national libraries

In 1974 the National Library of the Netherlands (KB) started the depository for Dutch publications in printed form, e.g. monographs, serials, newspapers, brochures, government publications and gray literature like Ph.D. theses.

The objective of the depository is to collect, describe, and store under optimal conditions and to give long-term access to the publications on site. A national depository collection is a last resort: when publications are no longer available on the market, the end user can always find a copy in the national library. In the Netherlands there is no law concerning the depositing of publications to the national depository as there is in many other countries. However, the depository for publications, working on a voluntary basis, has a high coverage of received documents. There is an ongoing discussion with the trade organization of the publishers concerning selection criteria and the conditions for depositing. Users can also find the description of the publications, via the World Wide Web, in the On-line Public Access Catalogue. It is only possible to get access to the publications in the national library itself, because it is a last resort library.

Many national libraries have accepted the rules for bibliographical description, published by the International Federation of Library Associations (IFLA). The IFLA has released the Universal Bibliographic Control Program (UBC), aiming to achieve bibliographical coverage worldwide. Each country should collect, describe, store and give access to the publications of their own territory and the bibliographical descriptions are published in the national bibliography. The criteria for selection and the types of material the countries include in the deposit are not the same everywhere. E.g. in Germany Die Deutsche Bibliothek collects audiovisual material, but in the Netherlands this task is carried out by the audiovisual archives of the Dutch public broadcasting company. There are several standards for bibliographical description published by IFLA, e.g. the International Standard for Bibliographic Description for monographs, ISBD (M) and serials ISBD (S), but there are also standards for antiquarian material, audiovisual material and recently for electronic resources, ISBD (ER). So, national libraries have a long tradition of selecting, acquiring, describing, storing and giving access to printed publications and sometimes other types of publications such as audiovisual material.

Many national libraries are now investigating how they can include electronic publications in the national bibliography and how they can store these documents and give access on site now and in the remote future. Often the law for the national deposit has to be changed to be able to include electronic publications. As mentioned previously, national libraries are faced with a wide range of judicial, organizational, financial and technological challenges, when they want to keep electronic publications accessible in the remote future.

Depository system for Dutch electronic publications

In the 1980's the KB every now and then received a book with a diskette or CD-rom enclosed. At the time the library did not have the facilities to view the elec-

tronic document and therefore a bibliographical description was made of the main work and the electronic publication was described as an enclosure. The diskette was shelved together with the book. In the 1990's it slowly became clear to the national libraries that electronic publications belong to the types of information they should collect and preserve. Therefore in 1994 the KB mentioned for the first time in the policy statement that it is an objective to collect electronic publications in the same way as printed publications. The topic was on the agenda with interested parties like the publishers and the Ministry of Education and Science[2]. The publishers are of course very much interested in terms of availability and copyright matters.

Following this first awareness the KB launched several actions to be able to organize the DSEP. A research project was carried out for the Steering Group of Innovation of Scientific Information Supply (IWI) in the period April 1996-December 1997. IWI is an initiative of the Dutch universities, the Royal Academy of the Sciences and the Dutch Organization for the funding of Research. Furthermore, several tests were carried out with (digitized versions of) scientific publications of the publishers Elsevier Science, Kluwer Academic Publishers and SDU Opmaat for official publications. In the mean time a study was commissioned by the European Commission concerning the depository collections of electronic publications. The project Cerberus investigated issues concerning authenticity and integrity of digital documents. BIBLINK investigated new ways of exchanging meta-data between publishers and KB.

Since 1995 the KB has followed the strategy of research, development, possible implementation, and joint research with organizations in Europe and the USA. Topics are selection criteria, acquisition methods, meta-data for discovery and technical information, unique identification numbers, methods for long-term preservation such as migration and emulation, storage on several types of carriers, methods for authenticity and integrity. Finally, IBM and the KB together are developing a system for preservation of electronic publications, which will be delivered in December 2002.

Related research and initiatives

The National Library of Australia's Preserving Access to Digital Information (PADI) initiative aims to provide mechanisms that will help to ensure that information in digital form is managed with appropriate consideration for preservation and future access.

Its objectives are to facilitate the development of strategies and guidelines for the preservation of access to digital information, to develop and maintain a web site for information and promotion purposes, to actively identify and promote relevant activities and to provide a forum for cross-sectoral cooperation on activities promoting the preservation of access to digital information.

The PADI web site is an excellent subject gateway to digital preservation

2 The Dutch Ministry of Education, Culture and Science.

resources. It has an associated discussion list padiforum-l for the exchange of news and ideas about digital preservation issues.

Worldwide people have realized that digital information is fragile. There are many initiatives and research projects investigating how electronic information can be kept accessible for end users in the remote future. Not only national libraries, but also archives, research institutes, banks and governmental organizations are facing the problem of how to keep digital information accessible in the remote future.
An important document for libraries was published by the Task Force on Archiving of Digital Information. The study was commissioned by the Commission on Preservation and Access and the Research Libraries Group.
Many national libraries, e.g. those of Australia, France, Finland, Germany and Norway, United Kingdom, United States are experimenting with the organization of a Depository for Electronic Publications.
Several European national libraries, together with publishers and computer companies, have carried out NEDLIB (Networked European Deposit Library) for Directorate General XIII of the European Commission. The results, e.g. a process model for a DSEP, a report describing an experiment in using emulation to preserve digital publications, guidelines for setting up a DSEP, etc. The documents are available via the Nedlib web site. The selected references at the end of this article give examples of relevant research and publications, also in the world of the archives.

Archiving the Internet

Several organizations are working on two methods for Internet archiving. The first is to make snapshots on a regular basis of a certain domain (worldwide, country-based). The second way is to select specific web sites or web publications. Several 'harvesters' are developed and they are available in the public domain or they are used by companies.
In 1998 the Internet Archive (IA) of Brewster Kahle in San Francisco made a copy of the 'entire' Internet and it continues to harvest large parts of the Internet. Recently the Way Back Machine was launched, which allows one to browse in the Archive. The largest national library in the world, the Library of Congress in Washington, has had an active, on-going partnership with the Internet Archive since 1998. Through a gift arrangement with the IA, the Library receives periodic gifts of segments of the Archive. The IA also provides frequent updates on a sustained basis of web sites in which the Library of Congress has a particular interest. Two examples of these special intensive-capture projects are the Election 2000 site and the September 11 project.
The objective of the IA is to archive large segments of the Internet for future users, but obviously the IA can't do everything that is required. It has estab-

lished a policy of not harvesting or making publicly available sites that either block the harvester or that ask not to have their sites captured. Therefore the Library of Congress is focusing on ways of obtaining any of these sites that they believe should be part of the historic record. They are doing this both through negotiating agreements outside of legal deposit provisions of US copyright law and through development of modifications to the legal deposit regulations that would require web site owners either to 'deposit' their sites or make clear that the Library of Congress' harvesting of those sites is not a violation of any copyright the creator might own in the site.

The PANDORA Archive of the National Library of Australia (NLA) is also exploring the documents available through the Way Back Machine. They have been looking at whether the IA makes any of their work redundant and whether it would make sense to join forces with the IA in a combined effort to ensure long-term availability of Australian on-line publications and web sites.

Experience in archiving web documents over the last five years has convinced PANDORA that where it is intended to preserve long-term access to a particular body of publications or information content, at this point in time given the current capabilities of harvesting software it is still necessary to create a separate archive, employing selection, quality control and preservation strategies and procedures, and what they have seen so far using the Way Back Machine has not changed their minds.

The Way Back Machine can deliver a giant step forward, but in the initial analysis of its success with a number of titles that are key to the PANDORA collection it was shown that it does not always capture the full information content. For example, internal links can be missing, parts of pages are missing as if the harvester has timed out half way through, and certain files types have not been captured. The harvesting software that PANDORA uses also sometimes archives incomplete copies, but staff then assesses the quality of each title gathered and remedy inconsistencies with the publishers' sites. Some titles require a lot of work, some require none. Another aspect of their work is to negotiate the permission to archive publications with publishers and make them accessible from the NLA site. This includes an increasing number of commercial publications, which would not be available to the Internet Archive harvester. The beauty of the Internet Archive is the breadth of its sweep and the full flavor of the Internet at any given time that it provides.

In the view of the NLA both approaches, the broad and the deep, will have value for future researchers, depending on the nature of the research. For the foreseeable future, selection and quality controlled archiving of publications judged to be of long-term research value will continue to be the first priority for PANDORA. In our experience, the purely automated approach is not yet sufficiently reliable to meet the information content needs of our users.

The Internet Archive and the American Library of Congress and the National Library of Australia are important examples of web archiving. But worldwide libraries and archives are working on the issues surrounding web archiving.

Summary

This document gives an overview of digital archiving from the perspective of national libraries. It concerns both archiving of off-line publications such as CD-rom and web archiving. The problem of the vulnerability of electronic information is not yet solved, but, worldwide, people are more aware of the issues concerning digital archiving.

References

- Hedstrom, M., S. Montgomery. (1998). Digital Preservation Needs and Requirements in RLG Member Institutions. http://www.rlg.org/preserv/digpres.html
- Jones, M., N. Beagrie. (2000). Preservation Management of Digital Materials Workbook. Pre-Publication Draft. October. http://www.jisc.ac.uk/dner/preservation/workbook/
- Lesk, M. (1997). How much Information Is there in the World? http://www.lesk.com/mlesk/ksg97/ksg.html
- Lyman, P., H.R. Varian. (2000). How much Information. http://www.sims.berkeley.edu/how-much-info/
- Reference Model for an Open Archival Information System (OAIS). (1998). Consultative Committee for Space Data Systems. http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html
- Rothenberg, J. (1996). Metadata to Support Data Quality and Longevity. http://computer.org/conferences/meta96/rothenberg_paper/ieee.data-quality.html
- Rothenberg, J. (1999). Avoiding Technological Quicksand, CLIR. http://www.clir.org/pubs/reports/rothenberg/contents.html
- Rothenberg, J. (1999). Ensuring the Longevity of Digital Information. http://www.kb.nl/kb/ict/dea/download/dig-info-paper.rothenberg.pdf

Web archives

- Australian National Library Web Archive. http://pandora.nla.gov.au/pandora
- Google's Web Cache Is a Form of a Short-Term Archive. http://www.google.com/search?q=cache:www.loc.gov
- Internet Archive and Way Back Machine. http://www.archive.org/
- Swedish Royal Library Web Archive. http://kulturarw3.kb.se/html/kulturarw3.eng.html

5

5.4 Text Mining

5.4.1 Understanding human language

Danny Lie[1]

Human language appears in two different forms: spoken and written. While this article mainly focuses on written language, the techniques that will be discussed also apply to spoken language. Besides form, there are two different modes in which human language can be used to communicate: by means of dialogue and by text. Both will be covered in this article, that focuses on the concepts. Section 6.2.20 will give a short overview of techniques used in text mining. See also Sections 2.2.1 and 2.3.10.

1 Ir D.H. Lie,
lie@carp-technologies.nl,
Carp technologies, Hengelo, The Netherlands
http://www.carp-technologies.nl

How can a computer understand human language?

Before we can answer this question, we fist have to discuss different levels of language understanding. There are three main levels of language understanding:

Level 1: Syntactic understanding.

Level 2: Semantic understanding.

Level 3: Pragmatic understanding.

For each level we will give an explanation and examples.

Syntactic understanding

For each message we want to convey in human language, we have to follow a 'communications protocol', which is defined in the grammar of the language. This grammar contains rules that state how words should be combined to construct valid sentences. For example, in English, an adjective should occur before a noun as in the phrase 'big dog'. Thus, a phrase like 'dog big' is incorrect according to this rule.

Inset 1: Glossary

proper noun	**name.**
article	**'the' or 'a'.**
adjective	**a word that modifies a noun by specifying an attribute.**
	Example: 'he hit the *red* ball'
noun phrase	**part of a sentence that describes a concept or object.**
	Example: '*he* hit *the red ball*'
verb phrase	**part of a sentence that describes an action.**
	Example: 'he *hit the red ball*'
subject	**subject of the main verb.**
	Example: '*John* gives Mary a book'
object	**first argument of the main verb.**
	Example: 'John gives *Mary* a book'
indirect object	**second argument of the main verb.**
	Example: 'John gives Mary *a book*'
agent	**the entity that performs the action.**
	Example: 'the red ball was hit by *him*'
recipient	**the entity that is the recipient of an action.**
	Example: 'he hit *the red ball*'
instrument	**the entity that is used as an instrument in an action.**
	Example: 'he hit the red ball *with his hand*'.

A sentence is a string of words, which have been combined according to the rules of the grammar. Thus, behind every sentence, there is an implicit structure

implied by the grammar. We call this structure the syntactic structure of a sentence. Consider for example the following sentence: 'John gives Mary a big book.' Each word in this sentence has a certain role: (see inset 1)

word	role
John	PROPER NOUN
gives	VERB
Mary	PROPER NOUN
a	ARTICLE
big	ADJECTIVE
book	NOUN

Furthermore, according to the grammar of the language, the words combine into different phrases, which also have certain roles:

words	phrase	role
John	NOUN	PHRASE SUBJECT
Mary	NOUN	PHRASE OBJECT
a big book	NOUN PHRASE	INDIRECT OBJECT
gives Mary a big book	VERB PHRASE	–

The analysis we have performed here is called a syntactic analysis, and can be visualized in a hierarchical structure (see Figure 1).

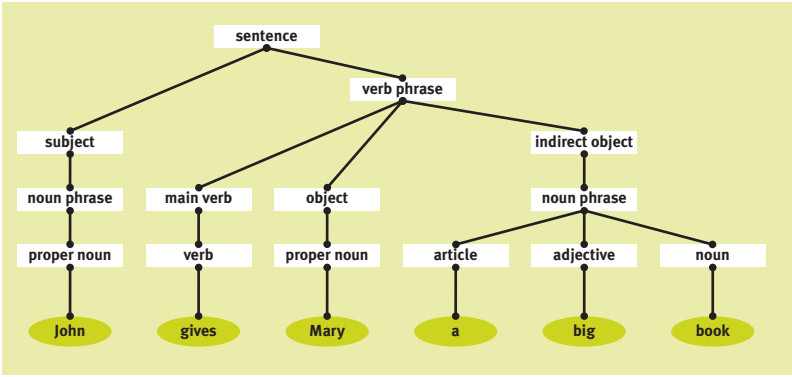


Figure 1
Syntactic structure of the sentence 'John gives Mary a big book'.

This structure is the syntactic structure of a sentence, and to be able to construct such a structure implies a syntactic understanding of the sentence. Note that the word 'book' can act as both a verb and a noun. But in this particular sentence, 'book' acts as a noun. This can only be determined after a syntactic analysis. A syntactic analysis can be performed by a computer by using a technology called natural language parsing.

Semantic understanding

Understanding the syntactical structure of a sentence does not imply an understanding of the meaning of a sentence (also called the semantics of a sentence). The next step is to determine the meaning of the phrase identified by the syntactic analysis and how they relate to each other. The following picture in Figure 2 shows this relation.

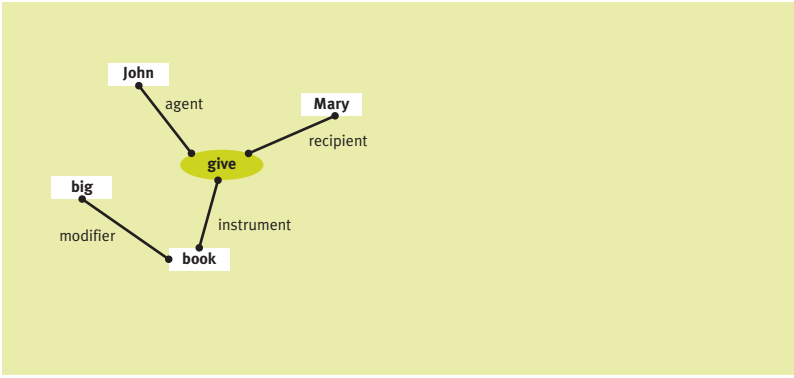


Figure 2
Semantic structure.

The meaning of this picture can be described as follows. There are three different objects: 'John', 'Mary' and 'book'. These objects participate in a 'give' relation in which 'John' is the agent, 'Mary' the recipient and 'book' the instrument. Furthermore, the 'book' object is modified by the attribute 'big', which indicates that its size is large.

The analysis we have just performed is called a semantic analysis, and the picture is a semantic structure. A semantic analysis can be performed by a computer using a technology called semantic analysis. Using other knowledge resources (such as ontologies[2]), the semantic structure can be decorated with more knowledge about the objects, deepening the level of semantic understanding (see Figure 3).

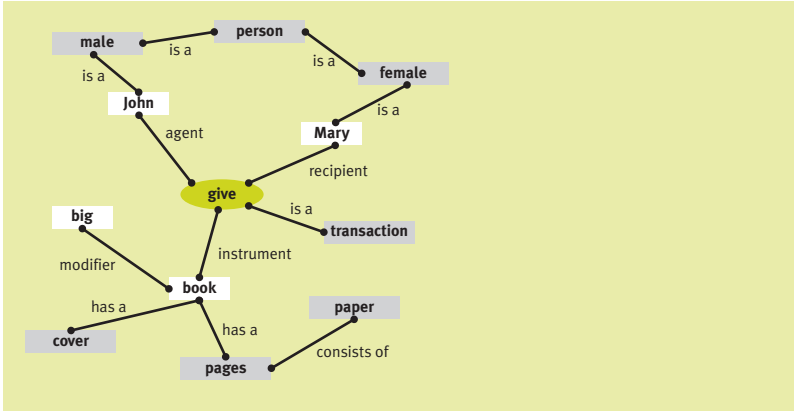


Figure 3
Semantic structure decorated with knowledge.

2 Explicit formal specifications of how to represent the objects, concepts and other entities that are assumed to exist in specific areas of interest and the relationships that hold among them.

The added information in the semantic structure can be described as follows: 'John' is a male and 'Mary' is a female, which are both persons. The 'give' relation is a transaction. Finally, a 'book' has a cover and pages which consist of paper.

A semantic analysis can be performed on a single sentence, but also on a text. Suppose our example sentence has been the first sentence of a text and that the next sentence is:

'She tears a page from it'.

Our existing semantic structure can be augmented with the semantic analysis of this sentence (see Figure 4).

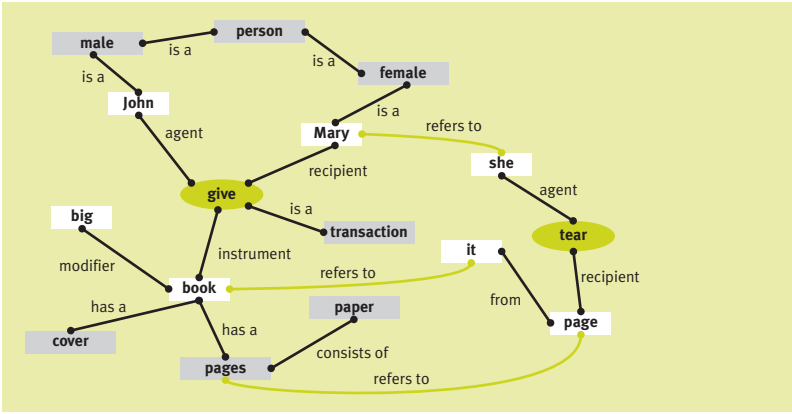


Figure 4
Augmented semantic structure.

One can imagine that the semantic structure of an entire text can be constructed by repeated augmenting of the semantic structure. The resulting 'web' of objects and relations represents the meaning of the text, and being able to construct it implies a semantic understanding of the text. Purely as an illustration, Figure 5 shows a computer generated semantic structure.

A very large and complicated web would be almost impossible for a human to untangle, but a computer can easily cope with it. Furthermore, a computer can perform all kinds of (mathematical) operations on the semantic structure, and in doing so, almost literally play with the meaning of a text!

Pragmatic understanding

For the analysis of a text, a semantic analysis is usually sufficient. However, in dialogues, people rarely say what they mean. Consider for example the sentence:

'Could you turn the light on?'

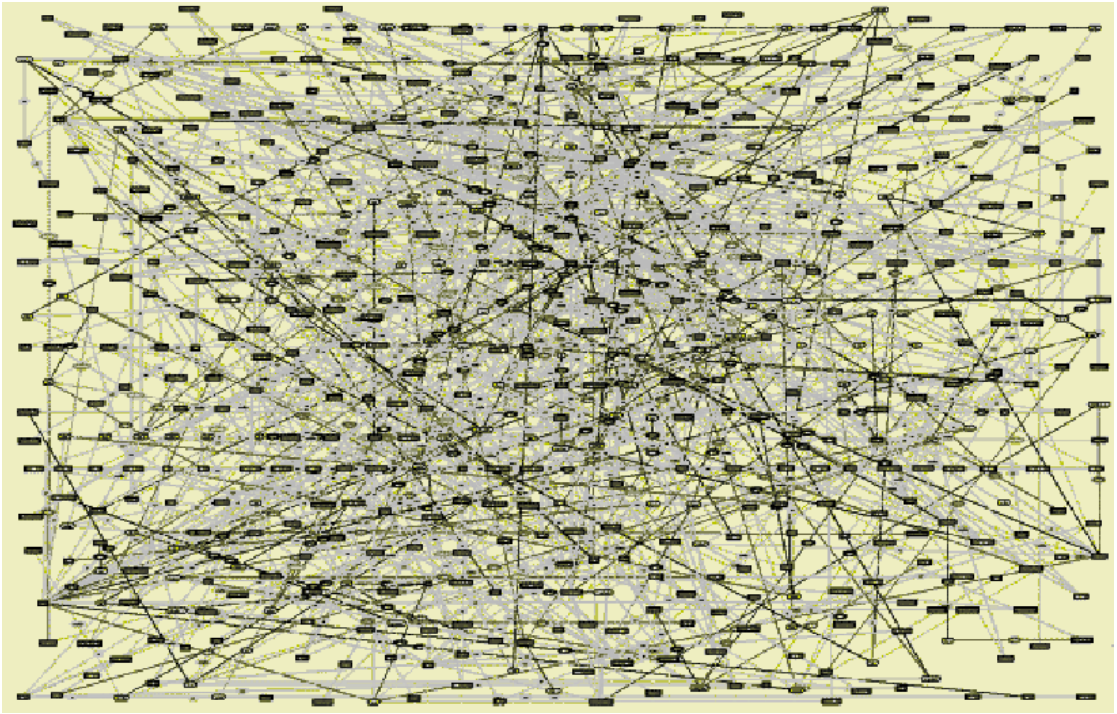


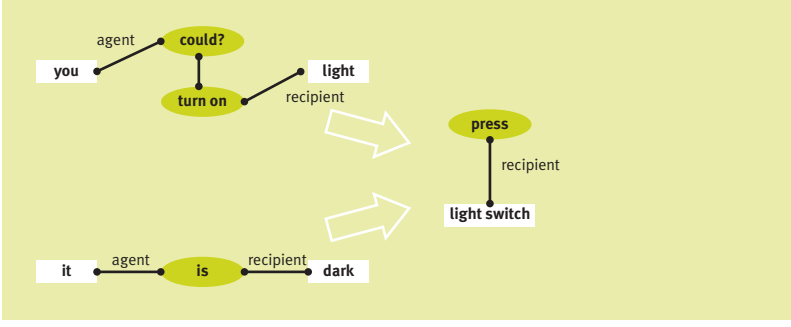
Figure 5
Computer generated semantic structure.

Obviously, the speaker means something like 'press the light switch', but the semantics of the sentence above is something like 'are you capable of achieving an increase in the amount of light?'. The former meaning is called the pragmatic meaning, and the latter the semantics of the sentence. Sometimes, the difference between the pragmatic meaning and the semantics is even more far fetched, as in the sentence:

'It is dark in here'.

Possibly, the speaker means exactly the same as in our first example: 'press the light switch'. To cope with these utterances, a computer has to make a mapping between the semantic analysis of a sentence and the pragmatic meaning (see Figure 6).

Figure 6
Mapping between semantics and pragmatic meaning.



For a computer to be able to make this mapping, it has to understand that a light switch is a device to turn on the light, and that an observation about the low light level could be a request to increase it. To be able to construct such mappings implies a pragmatic understanding of the utterance. The technology to construct these mappings is called pragmatic analysis.

The knowledge that is required to construct the mappings is called domain specific knowledge, or world knowledge. At the current level of technology, it is not yet possible to automatically construct domain specific knowledge. The technology to achieve this is still in its infancy, and is called machine learning. Therefore, domain specific knowledge has to be manually constructed for each specific domain. Consequently, it is not yet possible for a computer to converse about every conceivable topic. However, computers that converse about topics in a specific domain (for which the domain specific knowledge has been constructed) are definitely within our grasp!

Conclusion

Now let us return to the question: 'How can a computer understand human language?'. The answer to this question is now apparent: a computer can understand human language by applying a syntactic analysis, a semantic analysis, and optionally a pragmatic analysis.

5.4.2 Text mining

Marten Trautwein[3], *Quintus-Filius Grens*[4]

Introduction

History

The immense popularity of the Internet gave a boost to many technologies including text mining technology in the late nineties of the previous century. Initially, the Internet only required technology that distributed the information. When the amount of information increased, the need for processing technology became apparent. For example, in the early days a co-author would send his contribution for a text via a plain-text e-mail to the main author, who would incorporate that contribution. In comparison to telephone conversation or printed adjustments an enormous saving of time was achieved. Nowadays people are swamped with e-mail messages containing pictures, applications, sounds and the like. Only few messages are personally addressed and only few contain relevant information for the recipient.

Newcomers to the Internet still experience this shifting need from distribution to processing technology. The first day, they are enthusiastic they can share an MS-Word document with a friend via an e-mail message. Two weeks later they complain that they are spammed with many unsolicited e-mail advertisements[5]. The scenarios exemplified above do not only apply to the Internet, but also to intranets of organizations. People tend to drown in the quantity of electronically available data, which only partially contains valuable information.

Recent studies show that professional experts suffer from the problem of information overload. At first glance, the source of the problem seems to be the gross amount of available information. On consideration, however, the problem appears to be the inability to easily distinguish relevant and thus valuable from irrelevant and useless information.

In order to fruitfully discriminate information on relevance, the meaning of the information has to be understood. Various technologies are under development to extract the meaning of structured information, such as database tables [Witten, 1999] and unstructured information, for example pictures, movies and free texts [Witten, 1994]. Text mining technology attempts to reveal the meaning of unstructured textual information.

Related fields

As mentioned in the section above, text mining is one of many technologies that attempt to reveal the meaning of information. Text mining is considered to be a kind of machine learning technology [Mitchel, 1997] that operates on unstructured textual information. Text mining and data mining are closely related tech-

3 Dr M. Trautwein, marten.trautwein@ps.net, Perot Systems Nederland B.V., Syllogic Innovations, Amersfoort, The Netherlands

4 Drs Q. Grens, quintus.grens@ps.net, Perot Systems Nederland B.V., Syllogic Innovations, Amersfoort, The Netherlands, http://www.perotsystems.nl/, http://www.syllogic.com/

5 Phil Endecott writes "During June 1998 I recorded details of all the email I received. [..] 45% of my email was spam of some sort.", http://chezphil.org/spam/spam_stats.html

nologies. Data mining is a more mature and better known machine learning technology that operates on structured information, such as database entities. Machine learning technology (including among other things text mining and data mining) is capable of adapting to and learning from the circumstances in which it is applied. Thus, adaptive behavior takes into account that the meaning of information is not fixed, but subject to the context in which it occurs. For example, a cookery book on Italian cuisine not only contains information about recipes, but also about Italian eating culture.

We called the Internet the driving force behind many technologies. One of the most directly involved technologies is web technology. Web technology provides standards to construct infrastructures for sharing information. In other words, the web technology enables the distribution of information. Text mining, and data mining technology for that matter, utilize the web technology to reveal meaning in the information thus shared. An extension to web technology is (mobile) agent technology. Mobile agents can be viewed as small autonomous entities that travel over the infrastructure constructed with web technology. The mobile agents employ the web technology to collect information from the infrastructure. These mobile agents can be equipped with machine learning technology to improve their autonomous behavior. Mobile agents constructed in this way are often referred to as intelligent agents. A thorough discussion on agent technology is beyond the scope of this article (See [Tryllian, 2000; Taylor, 2001] and Van der Hoek, Section 2.2.2, Agents serving science).

Figure 1 below schematically depicts the relationships between the five above mentioned technologies.

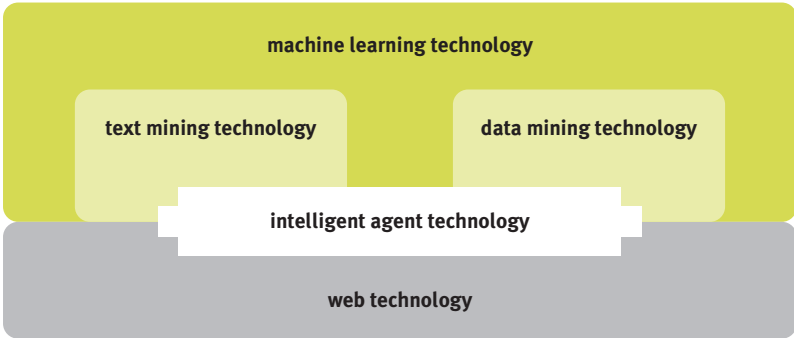


Figure 1
A schematic depiction of text mining technology and related fields. The machine learning technologies include text mining and data mining use, and build upon web technology. Intelligent agent technology, which heavily depends on web technology, uses the machine learning technologies.

Industrial applications

The most popular text mining application is document retrieval. Document retrieval is best known from the search sites, like Yahoo [Yahoo, 2001] and Google [Google, 2001], and the various search engines on individual web sites. Document retrieval, however, is broader than searching web sites, collections of text documents can also be searched (like the chapters in this book).

For example, a newspaper company may archive all its articles electronically.

A journalist working on some background story could use a document retrieval application to retrieve all previously published articles related to the background story's topic.

Less well-known text mining applications are name extraction, comparing, grouping, classifying or summarizing of textual information, such as reports, résumés, suspect profiles, news flashes, prospects, requests for proposals, call-center calls, etc. All these applications have a common purpose: sharing knowledge, reducing redundant information and increasing efficiency.

In a small project at the Dutch criminal investigation department text mining was used to compare, group and relate various reports of unsolved cases. The project yielded some new clues that the detectives had previously overlooked. These new clues enabled the detectives to solve some cases.

A French electric utility used text mining to analyze thousands of press-clips on the Internet regarding the agency's promotion of electric cars. Thus, they sampled public opinion without conducting interviews [IBM, 1997]. The same giant utility analyzed public data (such as newspaper, magazine articles) and customer claim letters to identify generic problems regarding customer needs. This categorization helped to adapt problems, uncover trends and develop pro-active strategies on customer needs [IBM, 1998].

At Professional Services Organizations (PSOs) new proposals to customers are often similar to proposals made in the past. Unfortunately, the company only benefits from the previous act, if the salesman writing the new proposal is familiar with the other proposal as well. Services companies frequently encounter this situation in which knowledge from the past can be reused in the future. Text mining technology is used to reveal the relevant knowledge from the past.

A Canadian provider of experience based diagnostic solutions to the aerospace industry used text mining technology to identify and assess the contents of electronic repositories of troubleshooting information. The text mining technology was used to perform content analysis and to create semantic networks that provided an overview of the important concepts of the repository [Megaputer Intelligence, 2001].

The biomedical literature consists of vast numbers of observations. The observations of interest to a single researcher are numerous and are distributed among many journals. The amount of information in the world's biomedical databases is growing so rapidly that the rate at which researchers can convert it to knowledge is falling behind. This uncovered knowledge could help to find cures for diseases like cancer and heart disease. The medical center of a university in the United States is aiming at using text mining tools and applying them to biomedical databases in order to reveal some of this uncovered knowledge and enable significant progress in biomedicine.

Current implementations

Most text mining applications are limited to index and retrieve functionality. A document collection is indexed, the index can be searched and documents are retrieved on the basis of the search criteria. Recent studies [Nielsen, 2001] show that users of document retrieval tools only use the basic functionality. Users hardly make advantage of the advanced (mathematically founded) options. In order to exploit the search functionality to its full extend, the tools should provide more intuitive options and display intelligence in helping the users. Thus, the prominent document retrieval tools add knowledge bases to increase the retrieval results. TextHub [Heijer, 1998] summarizes the retrieved documents to present a context[6]. TextAnalyst [Ananyan, 2001] and Yahoo [Yahoo, 2001] cluster the results to present related documents, Gartner-Search [Gartner, 2001] visualizes the category of the document. Only some specialized document retrieval tools do a coarse-gained analysis of the documents. For example, the search engine of CiteSeer, a digital library for scientific literature, extracts and indexes the documents' title, abstract and citation list [Lawrence, 1999].

Only few tools go beyond the document retrieval stage. The more advanced text mining applications translate documents, for instance Systransoft [Systransoft, 2001] and AltaVista [AltaVista, 2001], or summarize documents, such as Sinope (formerly known as Sumatra [Lie, 1998] and Inxight Summerizer SDK [Inxight Summerizer, 2001]. Both types of application serve the purpose to enable the user to quickly inspect the contents of documents. Especially in the case of long documents or documents in a foreign language these text mining applications save the user a lot of time.

One of the most widely used text mining tools [KDnuggets, 2000] is IBM's Intelligent Miner for Text, which is a toolkit rather than a single tool. The toolkit consists of various components that can be applied one after the other. Each component in isolation performs a very basic task, but, combined, sophisticated text mining applications can be created. For instance, a simple combination of components can forward generally addressed incoming e-mail messages to the group of interested persons automatically.

Future developments

Text mining

The introduction of the Internet has given text mining technology an enormous boost, and probably will remain the most important stimulator of future text mining technology. In addition, we expect stimulus from the office applications. Text mining technology will become an integrated part of office applications in the near future. Some text mining applications have clear correspondence with common tools in current office applications. For instance, any current word processing application should be able to check the spelling and preferably also the

6 TextHub is included as a demo on the CD-rom, mining the text of this book and additional articles.

grammar of the document. In the recent past the former and the latter functionality were considered highly innovative. Seen in this light, users of office applications should not be surprised, when the newest word processing applications display automatic summarization and translation functionality.
The next challenge in pure text mining technology will be interpreting a text. Summarizing and translating a text can be achieved to a large extend by means of statistical measures [Manning, 1999; Charniak, 1993], given a general dictionary and grammar. Both techniques focus largely on word level, a little at sentence level and hardly at discourse level.

The interpretation of texts requires a more thorough analysis of sentence structures. Current scientific research in computational linguistics concerns a field known as grammar induction [Bod, 1998; Daelemans, 1999; Vervoort, 2000; Adriaans, to appear]. This field tries in different ways to derive the grammar underlying a text from that same text. Grammar induction will, for instance, reveal a basic English grammar from a collection of children's books and a legal English grammar from a collection of insurance conditions.
These induced grammars are more fine-grained than the general grammars currently used, for instance in word processing applications. These fine-tuned grammars have the advantage that they describe the sentence structure more closely. Thus, these induced grammars will be better suited to indicate the discourse structure of the complete text as well.
The first step in interpreting texts is to determine the structure of the text. The next step is to analyze each segment in the structure separately. The analyses of the different segments will strengthen one another, when combined. These improved analyses will more precisely convey the message that the authors wanted to expose to the audience. For instance, segment a scientific article in abstract, introduction, examples, discourse and conclusions. The abstract will summarize the important statements of the article, which will take a prominent role in the rest of the article. The introduction will contain many forward looking statements, which will return in the conclusions in a recapitulated form. These statements will frequently occur in the main discourse as well, albeit not in the examples.

Knowledge management

Text mining technology in isolation will not conquer the world. The combination with other state of the art machine leaning technologies and especially data mining technology (see De Haas & Brandt in Section 6.4.3, Relational data mining) will impact the way people work to a larger extent. This combination of technologies will enable new knowledge management systems to act as intelligent (librarian) agents.[7]
Data mining technology will extract knowledge from structured data. Text mining

7 The ICT industry currently shifts towards component technology and open document formats. Small filter components extracting the structured and unstructured data from any type of documentation can be composed more easily. Henceforth, virtually all electronic documentation can be stored in such new knowledge management systems.

will reveal the knowledge in unstructured text data. The current scientific interest in analyses of pictures and movies will undoubtedly bring knowledge extraction from these types of unstructured data into sight as well.

Knowledge management is heading for the Trough of Disillusionment on the Future Technology Curve of Gartner Dataquest, because knowledge management is too labor dependent and intensive [Gartner Dataquest, 2001]. Gartner Dataquest estimates that knowledge management will be on the Plateau of Productivity within two to five years. This Plateau of Productivity will only be reached by applying machine learning technologies to the problem area of knowledge management. This will result in a next generation of intelligent adaptive knowledge management systems.

This next generation of knowledge management systems has two major improvements. First these knowledge management systems will interpret the documents stored automatically. As such the knowledge management systems will better understand the contents of these documents and thus more accurately satisfy user requests and expectations. Secondly, these knowledge management systems will learn the knowledge domain of the documents they contain automatically. Thus, these knowledge management systems will be able to increase their knowledge of the domain with the growing number of documents contained in the system. Moreover, the adaptive knowledge management system will be able to adjust its knowledge when documents from new domains are added to the system. The intelligent adaptive knowledge management systems will serve organizations in storing and retrieving documents, alerting users of interesting new knowledge and detecting potentially new associations.

The PSO industry has undergone significant changes. Multidisciplinarity is the current trend in this industry. The professionals should know more than their specialism, knowledge of related fields is required as well. Especially for multi-

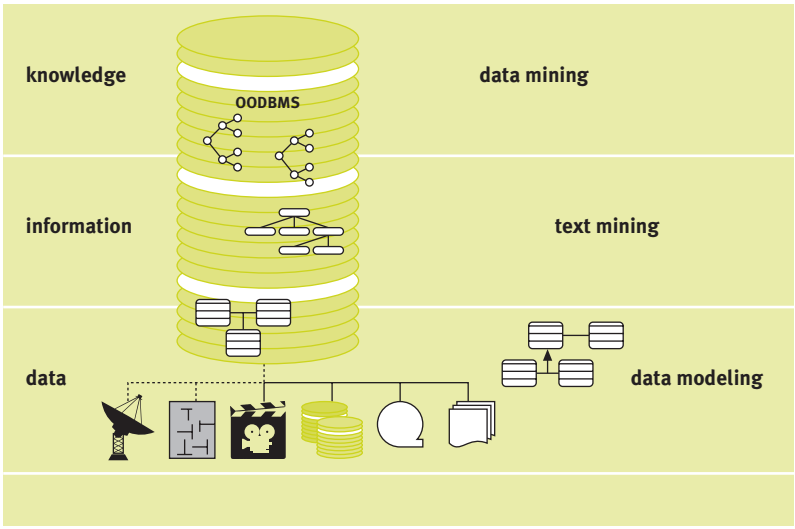


Figure 2
The schematic architecture of the intelligent adaptive knowledge management system which supports drilling down the levels, from knowledge, through the supporting information into the underlying data.

disciplinary and fast changing markets the next generation knowledge management systems will be of interest. In addition, customers view PSOs more as partners to compensate for the lack of expertise in house. This new view demands PSOs to provide customers long-term support. Professional Services Automation (PSA) [8] will help the PSOs with these changes [AberdeenGroup, 1999].

PSA promises to increase productivity of professional services personnel, reduce costs of engagements, and result in savings for the consumers. As a matter of fact, PSA solutions have the same potential for PSOs that Materials Requirement Planning (MRP) and Enterprise Resource Planning (ERP) have had in the manufacturing and distribution sector. The expected effect of PSA will be knowledge management improvements, improved productivity and ultimately, customer satisfaction.
The currently growing interest in the multidisciplinary field of biomedical informatics can already be seen in the direction of intelligent knowledge management systems. Huge numbers of articles have been published on the disciplines biology, medicine and genetics. In addition huge databases of structured data have been compiled in the past years. However, only a few scientists have worked on the interfaces between the disciplines and the few who have are overwhelmed with the overabundance of available data. Intelligent knowledge management systems will be able to make cross-links between the (structure and unstructured) data of the different disciplines. These cross-links will definitely assist the specialists from one field to become multidisciplinary experts. Figure 2 schematically depicts the architecture of the next generation of knowledge management systems.

The data modeling in Figure 2 accounts for the ordering and coupling of data, information and knowledge, abstracting away from the particular format of the raw data. At the middle level text mining techniques extract the concepts involved from texts and automatically construct a formal representation of the domain knowledge. At the upper level data mining techniques generate new knowledge from the data using the information stored in the middle layer. Since the data mining techniques use this information layer, the domain knowledge is taken into account, yielding accurate domain specific new knowledge.

Investments

Gartner [Gartner Group, 2000] estimates the total customer expenditures for text mining in North-America and Western Europe to increase from $200 million in 1999 to over $8 billion in 2003. This growth will depend on the functionality of the text mining applications as well as the capabilities of the text mining application integrators.

8 Professional Services Automation (PSA) is also known as business process automation, services industry applications, services process optimisation and services relationship management.

Text mining applications will only be effective, when domain knowledge of the application domain can be incorporated. Third party system integrators will have to implement and customize the text mining solutions, since no shrink-wrapped solutions will be available for the near future. As such, early text mining projects will become multi-million dollar efforts with high visibility in the user organization.

The authors were personally involved in a few knowledge management projects of knowledge intensive customers. These customers knew their organization was using its knowledge inefficiently: reinventing the wheel regularly and duplicating work instead of reusing work. The total costs of their activities, e.g. conducting studies, responding to a request for proposals, were known. However, no index numbers were collected on the cost of the inefficient use of knowledge in the organization. Therefore, the return on investments for these projects was measured in terms of saved effort. If the customers could only prevent one unsuccessful study or gain one additional contract through a request for proposals, the investments would be covered.

The implementation of a PSA solution requires at least an increase in the infrastructure of up to 15%. In exchange, the productivity may rise up to 40% and cost may be reduced in billing and project management. The expected business on PSA will increase with 70% for the coming years (from $300 million in 2000 and $500 million in 2001). PSA is expected to have a pay back time of one year, which may become 2 to 6 months in the optimal case [Automatisering Gids, 2001].

Conclusions

The amount of structured and unstructured information shared between people has increased drastically, since the introduction of the Internet. Text mining technology assists to find the valuable information by revealing the meaning of unstructured textual information.

Most current text mining applications are limited to index and retrieve functionality. Only few tools currently go beyond the document retrieval stage. The next challenge in text mining technology will be interpreting a text, and conveying the message the authors wanted to expose.

The combination of current and emerging machine leaning technologies, especially text mining technology and data mining technology, will impact the way people work. Combining these technologies with general machine learning technology will yield a next generation of intelligent adaptive knowledge management systems. These knowledge management systems will be able to grow their knowledge of the domain with the growing number of documents contained in the system. Moreover, the adaptive knowledge management system will be able to adjust its knowledge, when documents from new domains are added to the system. Especially for multidisciplinary and fast changing markets, such as the professional services organization industry, the next generation

knowledge management systems will be of interest.
PSA, integrating knowledge management, is expected to have a pay back time for PSOs of one year. The early text mining projects will become multi-million dollar efforts with high visibility in the user organization.

References

- AberdeenGroup. (1999). Professional Services Automation: Increasing Efficiencies and Profitability in Professional Services Organizations. AberdeenGroup, Boston. http://www.aberdeen.com/ab%5Fcompany/hottopics/psa/psa_execsum.pdf
- Adriaans, P.W., M.H. Trautwein, M.R. Vervoort. (2000). Towards High Speed Grammar Induction on Large Text Corpora. Proceedings of SOFSEM 2000. (to Appear)
- AltaVista. AltaVista - The Search Company. http://www.altavista.com/
- Ananyan, S., A. Kharlamov. Automated Analysis of Natural Language Texts. Megaputer Whitepaper. http://www.megaputer.com/tech/wp/tm.php3
- Automatisering Gids. (2001). Met stroomlijning van adviesprocessen is geld te verdienen. Automatisering Gids. http://www.automatiseringgids.nl/
- Bod, R. (1998). Beyond Grammar: An Experience-Base Theory of Language. CSLI Lecture Notes Number 88. CSLI Publications, Stanford
- Charniak, E. (1993). Statistical Language Learning. The MIT Press, Cambridge
- CiteSeer. (2001). ResearchIndex [NEC Research Institute; Steve Lawrence, Kurt Bollacker, Lee Giles; Computer Science]. http://citeseer.nj.nec.com/
- Daelemans, W., A. van den Bosch, J. Zavrel. (1999). Forgetting Exceptions is Harmful in Language Learning. Machine Learning **34**:11-43
- Gartner. (2001). Gartner.com. http://www.gartner.com/
- Gartner Dataquest. (2001). Directions and Trends in the IT Professional Services Market. Gartner Group Presentation. 22 January
- Gartner Group. (2000). Text Mining. GartnerGroup White Paper. http://www.xanalys.com/intelligence_tools/products/text_mining_text.html
- Google. Google. http://www.google.com/
- Heijer, E. den, K. Blok, M. Trautwein. (1998). Document Clustering in a Information Retrieval Environment: TextHub. Proceedings of the Eighth Belgian-Dutch Conference on Machine Learning. pp11–19
- IBM. Speed Reading the Internet. (1997). IBM THINK magazine **2**. http://www.ibm.com/software/data/iminer/fortext/thinkArticle.html
- IBM. (1998). The Power of Enlightenment. IBM Global Business Intelligence Solutions. http://www.ibm.com/software/data/bi/pdf/edf.pdf
- Inxight Summerizer. Eliminate the Guesswork from Online Searches. http://www.inxight.com/pdfs/products/summarizer_sdk_ds.pdf

- KDnuggets News. (2000). Which Text-Mining Packages you Have Used? Polls. http://www.kdnuggets.com/polls/text-mining.htm
- Lawrence, S., K. Bollacker, C.L. Giles. (1999). Indexing and Retrieval of Scientific Literature. Proceedings of CIKM 99. pp139-146. http://www.neci.nec.com/~lawrence/papers/cs-cikm99/cs-cikm99.pdf
- Lie, D.H. (1998). Sumatra: A System for Automatic Summary Generation. Proceedings Twente Workshop on Language Technology **14**. http://www.carp-technologies.nl/SumatraTWLT14paper/SumatraTWLT14.html
- Manning, C.D., H. Schútze. (1999). Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge
- Megaputer Intelligence. CaseBank Textmining Case Study in Banking. http://www.megaputer.com/company/cases/casebank.php3
- Mitchel, T.M. (1997). Machine Learning. The McGraw Hill, New York
- Nielsen, J. Search: Visible and Simple. Alertbox. http://www.useit.com/alertbox/20010513.html
- Systransoft. Systran. http://www.systransoft.com/
- Taylor, A.A., S. Garone. (2001). Agents on the Move. An IDC White Paper. http://www.tryllian.com/sub_downl/2915_rev2.pdf
- Tryllian. (2000). Mobile Agents: Going beyond the Web. A Commercial White Paper. Version 1.0
- Vervoort, M.R. (2000). Games, Walks and Grammars: Problems I've Worked on. ILLC Dissertation Series DS-2000-03. Institute for Logic, Language and Computation, University of Amsterdam. http://www.illc.uva.nl/Publications/Dissertations/DS-2000-03.text.pdf
- Witten, I.H., A. Moffat, T.B. Bell. (1994). Managing Gigabytes: Compressing and Indexing Documents and Images. Van Nostrad Reinhold, New York
- Witten, I.H., E. Frank. (1999). Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufman Publishers
- Yahoo. Yahoo! http://www.yahoo.com/

5

5·5 Multimedia Mining

5.5.1 The infant days of multimedia data mining

Osmar Zaïane[1]

Where do we stand?

The low price of storage devices has accelerated and significantly increased the storage of what is conventionally called multimedia data. For instance, a significant amount of data such as surveillance camera video streams, satellite and aerial pictures, medical images, music and sound files, etc. are now commonly stored in digital form. Efficiently managing this type of data and successfully retrieving essential items when needed is a crucial issue. This need has attracted the attention of many researchers in database management systems and information retrieval, and significant achievement has already been realized in image indexing and content-based image retrieval. However, when amassing large collections of data, accurately retrieving some items from the collection which are relevant to a certain need is not the only issue. The task of discovering general trends and hidden patterns from within the large collection becomes essential for decision-makers. This is also true for large collections of multimedia objects in many applications and domains. This activity is a new one and is at the confluence of many research fields.

1 Dr O.R. Zaïane, zaiane@cs.ualberta.ca, University of Alberta, Department of Computing Science, Alberta, USA, http://www.cs.ualberta.ca/~zaiane

What is multimedia data mining?

Multimedia data mining, in principle, refers to the set of methods specifically devised for the extraction of hidden knowledge from within assortments of media, including text, images, sounds, videos, maps, etc. It is an integral part of the research field of knowledge discovery and data mining (KDD) which aims at discovering interesting, implicit and previously unknown patterns from large collections of data. While one of the early works in this field of knowledge discovery for databases pertains to the recognition and classification of objects in satellite images [Fayyad, 1993; Fayyad 1996], data mining from multimedia is nonetheless evolving at a slow pace. Multimedia data mining is still in its infancy. Very few significant achievements have been realized, when it comes to extracting useful knowledge from inside visual or audible media. In fact the discipline is neither prevalent nor clearly defined. The scope of multimedia data mining is still being debated. Even experts do not jointly agree what this new research field encompasses. While the purpose of data mining, the discovery of implicit patterns in large collections of data is clear, many consider content-based retrieval from image or video collections, for example, to be a data mining task. Moreover the range of media to be considered in multimedia data mining is also still unclear. Whether multimedia data mining is an extension of data mining involving all sorts of data or a particular case of data mining considering still images, video and sound only, is still unresolved. It is well understood or agreed upon that multimedia has come to mean, or refers to, images, text, graphics, sound, motion pictures, video, and even animation, games, maps and virtual reality. In other words, it could be any combination of data in digital format. Whether a document involves several media (synchronized or not) or a single medium, it is deemed multimedia. However, in the multimedia data mining arena, multimedia is often confined to still images, audio and video, and occasionally includes text and spatio-temporal data. Text mining and spatial data mining, or spatio-temporal data mining, however, are well-established research disciplines, and tend to demarcate themselves from multimedia. In particular text mining is a very active research field, often involving natural language processing that concentrates on semi-structured or unstructured text documents for text categorization, clustering, or the extraction of associations between terms and phrases, also known as KDT (knowledge discovery from text) [Feldman, 1995]. Spatial data mining on the other hand offers techniques for topological knowledge extraction that can be used in image content mining. Spatial data, also known as geographic data, consists of spatial objects with geometric and topological properties. These intrinsic properties revealed in geographic data are often sought for in image data and extracted whenever image processing makes it possible. The methods used in spatial data mining to discover knowledge from topological relationships between objects in maps are analogous to the methods used for discovering knowledge from within images.

The First International Workshop on Multimedia Data Mining [Zaïane, 2000b] was held in conjunction with the ACM Knowledge Discovery and Data Mining conference in Boston in August 2000. Based on the papers presented at the workshop and the discussion held, it is clear that the extent of multimedia data mining is not widely agreed upon [Simoff, 2001]. Nevertheless, while the workshop included geo-spatial data mining, the compromise is that multimedia data mining unquestionably includes mining from still images, mining from video, and mining from sound.

What is not multimedia mining

The process of mining is poorly understood and often leads to misconceptions in the case of multimedia mining. Given a collection of multimedia objects, multimedia data mining consists of extracting some implicit and previously unknown patterns from the collection. The patterns can come from the ensemble of the objects or from within the content of these objects. The patterns to be extracted are said to be implicit, because they are not explicitly included in the collection, but implied from the content or the relationship between the objects in the collection. This is the fundamental difference with the retrieval of objects from the collection known as Information Retrieval (IR), because information retrieval finds explicit and known objects in the collection and not patterns. While the major input to these two processes is essentially the same, the collection of multimedia objects, the output is profoundly divergent. Information retrieval discovers relevant objects in the collection of multimedia objects, and multimedia data mining discovers patterns (i.e. knowledge) from the same collection. Even among scholars this misconception still persists. However, content-based information retrieval is not multimedia data mining. Multimedia data mining encompasses the typical canonical data mining tasks such as generalization, classification, clustering, association mining, sequence analysis, outlier detection, etc., and often necessitates difficult and tricky data preprocessing in order to extract pertinent content features that are used in the mining task at hand. Information retrieval necessitates the same tedious content-feature extraction from multimedia objects for indexing purposes. This feature extraction can often be considered as data mining. Thus, multimedia data mining can play an interesting role in information retrieval, when categorizing multimedia objects and creating indexing schemes. Indeed data mining can be a means of discovering interesting features used in indexing multimedia objects, be it images, sound or video streams.

Another misconception is the link between multimedia mining and visual computing and processing. Visual computing includes computer visualization and the use of artificial intelligence in visual applications. This could indeed be related to the tasks pertaining to multimedia data mining, in particular for

images and video, however, visual computing also includes digital image generation, visual animation, etc. that are not associated with the knowledge discovery process.

Challenges in multimedia mining

While knowledge discovery from data is a well-established field and many data mining techniques are becoming mainstream [Han, 2000], there are many challenges still to overcome in particular for multimedia data mining. Outlined below are some of the major issues and challenges that face the multimedia data mining community as a whole:

Complexity of data: Conventional numerical and categorical data is well structured and organized assisting to some extent in the knowledge discovery process. However, multimedia data is often unstructured, difficult to interpret and stored in a variety of different formats making data mining of such media more problematic. Formats are also different from one medium to another and are often proprietary. Moreover, in multimedia applications, many media are combined and carefully synchronized adding to the difficulty in mining such data.

Scalability: The data mining community advocates techniques that can handle very large data sets. Commendable data mining techniques are techniques that can handles millions of transactions and more. Multimedia is known to be usually very large and to require significant storage space in comparison to conventional numerical data. Whether it is video, images or audio, collections are tremendously large, and thus are well suited for data mining. However, what is thought of as very large in conventional databases is in the order of some gigabytes. Multimedia databases on the other hand can easily reach hundreds of gigabytes and even terabytes in size. Scalable tools and algorithms for preprocessing and mining multimedia that can manage such extremely large data in a reasonable time are yet to be developed. Massively parallel and high performance computing should help in this perspective for both multimedia data preprocessing and multimedia data mining.

Regularity: Multimedia repositories can contain large collections of multimedia objects of all sorts. For example, the World Wide Web contains all kinds of images, videos and sound files with a variety of subjects. Mining such collections could yield the discovery of rules, but the validity or the interpretability of such rules is questionable. In other words, for data mining to be beneficial and discovering realistic and valuable rules from multimedia, the objects in the multimedia collection to be mined should be homogeneously and semantically grouped and belonging to the same application domain. For example, patterns extracted from a collection of satellite pictures mixed with images of human faces, and fish in the deep sea would not be very useful. However, discovering patterns in a collection of medical images all presenting brain scans can produce

rules which are medically sound. Categorizing multimedia objects for rigorous and reliable data mining is a difficult task. This is particularly true, when it comes to sound streams. Sound is often polyphonic and necessitates segmentation in monophonic channels for further processing.

Privacy: Privacy has always been an important issue in data gathering and access. Data mining made this issue even more significant, when it comes to adequately interpreting the knowledge discovered and addressing the legal and ethical question of invasion of privacy. Multimedia data mining with applications such as surveillance cameras and medical imaging propels the problem of privacy a step further.

Data inaccessibility: A paradox in multimedia data mining research is data inaccessibility, even though multimedia data are said to be abundant and ubiquitous. Gathering large quantities of semantically related multimedia data for research purposes or even industrial applications is not an easy task. This becomes a major problem in particular for research, since relying on small data sets or on synthetically generated data is often deceptive and the techniques devised become inappropriate in real life settings. Data acquisition and selection is fundamental in the knowledge discovery process. The reasons for such inaccessibility are manifold. The data can be distributed by operation strategies, storage availability, transmission costs or simply remote sensing capabilities. Bureaucratic problems, rigid governmental and administrative rules, and privacy issues can also hinder accessibility, for example in the case of medical images and satellite images.

Insufficient training: Knowledge discovery and data mining are at the confluence of many disciplines: machine learning, artificial intelligence, databases, statistics, high performance computing, visualization, etc. Multimedia data mining adds to the equation image processing, vision and signal processing [Zaïane, 2000a]. Interdisciplinary skills are required to process and cope with multimedia data. Such polyvalence in individuals and teams are hard to obtain.

Insufficient support: Most versatile data mining tools are for generic numeric data, usually in flat files or relational databases and are rarely adaptable for the purpose of multimedia. There is no significant development of specific algorithms for multimedia data mining applications and tools. Multimedia data mining tools should be flexible to adapt to domain knowledge and domain applications, even if they are restricted to a specific medium such as image or video. Tools should also be integrated to allow ad hoc data mining with different knowledge discovery techniques. Multimedia data mining relies heavily on fields such as vision and signal processing for data preprocessing and feature extraction. These fields also lack adequate tool support.

The current trends

Multimedia data mining is defined in [Zaïane, 1998a] as being the mining of high-level multimedia information and knowledge from large multimedia databases. This consists of discovering patterns from descriptions of multimedia objects, either from within the objects or descriptions accompanying the objects. Thus, there are two types of multimedia data mining practices: content-based multimedia data mining, which discovers patterns primarily from the content of multimedia objects, and description-based multimedia data mining, which concentrates on external descriptors [Zaïane, 2000]. Often these practices are combined, for example the system described in [Zaïane, 1998b] is a system that discovers characteristics, classes, and associations from a data cube that aggregates data from image descriptors. Most of these descriptors are peripheral to the images such as size and keywords, but some others pertain to dominant colors and dominant edge orientations in the images. Such systems are useful for visual asset management, but are limited in their knowledge discovery capabilities. The patterns discovered, moreover, could be questionable as a result of the heterogeneity of the image collections dealt with. Early multimedia data mining systems aimed at discovering high level, general purpose patterns and some others used multimedia as a support for data mining results analysis and visualization such as [Bhandari, 1997; Noirhomme-Fraiture, 2000]. The new trend, however, is in a specific application domain such as categorizing in medical imaging, tracking objects in video, detecting narrative structures in news broadcasts, etc. as we shall see below.

To our knowledge, there is no system today that discovers knowledge from general repositories of multimedia, but typically, multimedia data mining systems would focus on specific media, whether it be images, sound or video. Even systems for mining newscast such as [Shearer, 2000] often convert speech into text and mine text and video streams rather than sound per se and video.

Mining still pictures

While the input of a content-based image retrieval system is a set of visual features from a collection of images and a sample image or query describing characteristics of images to be retrieved as output, the input of an image mining system is the same set of visual features from a collection of images along with a specific data mining task to apply, and the output is knowledge usually in the form of rules. As described above, the data mining of images can be either content-based or description-based. In both cases, the knowledge discovery process uses features describing the images to carry out the mining task. As discussed in the previous paragraphs, in content-based image mining the (low-level) features are usually (technical) visual descriptors such as colors, textures, shapes, objects, or spatial topological relationships like overlap, inclusion, etc.

The description-based mining of images focuses on external descriptors such as size, format, creation date, associated keywords, etc. In many applications, however, these two types are combined to take advantage of both descriptors. For instance, in a knowledge discovery process involving medical images, visual features such as shapes and textures will be used along with diagnoses attached to the images. Moreover, when mining from images, image features are often combined with other relevant data that is not necessarily in or about the image per se. Considering the example with medical images, characteristics attached to the patients files such as gender, age, medical history, etc., could be taken into account in the knowledge discovery process.

There are different canonical data mining tasks that can be pursued in knowledge discovery for images: characterization, supervised classification, unsupervised classification and association rule mining.

Characterization

Characterization of images consists of summarizing image descriptors to provide a general portrayal of the image collection. This summarization usually proceeds by a generalization along some concept hierarchies defined for each descriptor of the images, a method primarily based on attribute-oriented induction [Han, 1993]. This provides an on-line analytical processing (OLAP) capability for drilling down and rolling up in a multidimensional data cube such as in the MultiMedia-Miner system described in [Zaïane, 1998b].

Classification

Classification of images (also known as image categorization) or classification of image content is a very important data mining task for many applications that handle images for knowledge extraction. Perhaps one of the most influential papers in the early days of data mining [Fayyad, 1993] described a classification system that categorized high-resolution radar images of the surface of Venus transmitted by the Magellan spacecraft in order to distinguish and identify volcanoes on the surface of the planet. One of the most important and fundamental lessons learned from such a system is the heavy cost of preprocessing in the case of multimedia data mining. Images have to be segmented in order to identify areas of the images relevant to the task at hand, and to transform the extracted data into a representation on which existing algorithms can operate. In the case of [Fayyad, 1993], a known decision tree induction algorithm, ID-3, was used. In [Bamford, 1999], a Hidden Markov Model (see 6.2.10) was used to discriminate between easy and hard observations of cell nuclei in the case of detection of cancer of the cervix.
Classification in the context of multimedia mining is not only used for discriminating part of images such as in the previous cases, but also to categorize entire

images. In particular in the domain of medical imaging, techniques such as neural networks and decision tree induction are used to build classifiers for mammography, retinal photography, and various medical scans.

Clustering

Clustering, also known as unsupervised classification, aims at grouping similar objects together. While data mining and machine learning researchers are very active in the field of clustering, not many applications have used clustering on images. A known clustering Algorithm, BIRCH, has been used to achieve segmentation of pixels in near-infrared images in order to create filters on pictures in a soil science application [Zhang, 1996]. Clustering is also used in the case of satellite images and low aerial photographs such as in [Blackard, 1999] investigating forest cover type. Such studies are very useful for mapping agricultural areas, estimation of snow (water) quantity, detection of growth of algae, detection of oil spills, surveillance, etc.

Association rule mining

The discovery of associations between items in a data collection is one of the most important and studied tasks in data mining. It consists of finding items that frequently occur together in a collection. This is usually applied in transactional databases such as in market basket analysis. Association rule mining has been applied in the context of images where images are transformed into transactions, each transaction representing the objects (or entities) in a image. In [Ordonez, 1999] for instance, the a priori algorithm was used to discover simple associations between elliptical regions that are coherent in color and texture in an image, called blobs. Another study of associations between feature localizations in images has investigated the reoccurrence of visual features and their associations as well as their spatial relationships [Zaïane, 2000b]. These topological associations of visual features in images are of major interest in the analysis of brain scans, for example, and a means for potential discovery of relevant patterns in very large collections of scans that human beings cannot browse and analyze manually in a reasonable time.

Mining video

Digital video cameras are becoming relatively cheap and ubiquitous. This has increased the possibility of amassing huge collections of video which are already very large. Surveillance cameras for example are almost ever-present. It is not realistic to browse the video streams manually or go over all available video streams in search of some events. The information retrieval approach attempts to find video sequences that correspond to some specified descriptors, such as objects in the scene or occurrence of a described incident. However, the retrieval usually operates on manually annotated video. The data

mining approach usually alleviates the need for manual intervention and either attempts to discover general patterns or extract useful features that could be used in video querying, for example automatically annotating video sequences. Automatic annotation of digital video could yield significant advancement in video management and retrieval, since the manual annotation is time-consuming and expensive. In an effort to summarize news broadcast for retrieval and browsing, video segments in news broadcasts have been classified and automatically labeled: anchor shots, footage with voice over or sound bite, after segmentation and narrative structure detection tools have been applied on the data stream [Shearer, 2000]. However, the sound channel is not used, but is converted to written text with speech-to-text tools. Another application that exclusively uses video streams to extract advantageous knowledge from within video is a surveillance application that tracks unusual behaviors of individual movements in a crowd [Tucakov, 1998]. The system uses stereo cameras to produce distance measures in 3 dimensions and tracks movements of objects in path trajectories where each path trajectory is a spatio-temporal activity of individual objects in a 2-dimensional space. Using a method for outlier detection in a set of path trajectories, the system is capable of identifying suspicious behavior from videos taken by surveillance cameras. Modeling movements in internal representation for mining patterns and tracking objects in real time is used in many industrial and military applications. For example, cameras mounted on automobiles can identify and track pedestrians or other objects and alert the driver to hazards. The system described in [Papageorgiou, 1998] learns to identify pedestrians and cars using a wavelet representation of different observations and a vector machine classifier. Video mining is discussed elaborately in Section 5.5.4, Data mining for video retrieval.

Mining sound

While dictation software, voice command applications for desktops, voice mail and text-to-speech converters are becoming mainstream, sound coding, analysis and indexing are still at a very early stage. Sound mining is even less advanced. Often data mining from sound is not seen as a goal, but as a method for processing sound data. For example data mining or machine learning techniques have been used to filter sound sources by classifying sound channels. Neural networks techniques were used to enhance old audio recordings by eliminating noise [Czyzewski, 1994]. The main investigated knowledge discovery task in the context of sound or speech is incontestably audio signal classification. However, data mining from sound remains extremely difficult due to the necessary tricky preprocessing of the audio signal. The auditory scene is usually polyphonic (i.e. different sounds combined into a single signal) making it extremely difficult to distinguish between different overlapping sound steams. Because of the complexity of sound stream segmentation, many applications

consider only monophonic signals. As stated above, in data mining applications involving speech, speech-to-text converters are often used to take advantage of existing text mining and text analysis tools. Yet, there are interesting developments in sound classifiers to distinguish between speech and music, categorize human accents, identifying speaker gender, etc. This trend in research on building sound classifiers comes from the conviction that sound classifiers could potentially help us enhance our perception of sounds. New research in clustering of sound segments aims at determining groupings in human speech segments, or classification schemes for indexing sound segments [Matichuk, 2000]. These indexing schemes could help in automatic generation of speech with accurate prosodic patterns giving synthetic speech with more natural characteristics. The study is also investigating the use of the sound segment clustering techniques in new speech compression approaches. This corroborates the statement that multimedia data mining is often used as a means rather than just a goal. A special case is the analysis of music, which is discussed in Section 5.5.2, Musical audio mining.

Considering the future

We are witnessing the dawn of a new application and research field in knowledge discovery from data: multimedia data mining. The current applications are basic and the techniques used are still rudimentary. However, mining from multimedia data will soon become mainstream. The current trends are to apply existing knowledge discovery approaches. Adapting existing algorithms, however, does not always yield the sought for results, and significant changes to the algorithms are often necessary. In the near future we will see a drift towards new multimedia mining dedicated algorithms, algorithms specifically designed for the high-dimensional, rich and complex multimedia data. Until new approaches purposely developed for multimedia are conceived, it is difficult to extend the limits of multimedia data mining.

Multimedia data mining is a convergence of many fields and expertise. Whether mining images, sound or video, a significant proficiency in a variety of practices is required for data preprocessing. Image processing, vision and signal processing are modus operandi for multimedia data mining and will become more involved with dedicated approaches to knowledge discovery in multimedia data. Many research communities have been working on related problems such as image interpretation, media understanding, vision, etc. for many years. However, most of the work has been concentrated on small data sets. Multimedia data mining is more devoted to very large collections and scalability is a very important issue. What will give a serious boost and high momentum to the research in multimedia data mining is the awareness of the research communities for the new needs, and the close collaboration between researchers

with different expertise, insight and talent. This collaboration will bridge the chasm between early prototyping research and real mainstream applications of multimedia data mining.

In the very near future, we will see multimedia data mining tools as conventional applications in cars, in homes, and even with wearables (i.e. computer powered cameras, built-in garments). Cameras mounted on computer displays could identify user emotions and interpret needs. Identifying and recognizing objects in real time would become a common task for satellite pictures and mobile digital cameras. Cameras mounted on mobile carriers, such as cars or even humans, will have enough computing power to become robust autonomous mobile miners to help users recognize and interpret the environment in which they proceed. For instance, such devices will become an integral part of any modern car, an essential driver's aid to track moving objects in the environment or monitor the occupants for fatigue or distraction.

There are many applications of multimedia data mining, and many others are still to come, from medical imaging to weather forecasting and from video-taped sporting events analysis to speech compression and generation. The applications of multimedia data mining are only bounded by our imagination.

References

- Bhandari, I., E. Colet, J. Parker, Z. Pines, P. Pratap. (1997). Advanced Scout: Data Mining and Knowledge Discovery in NBA Data. Data Mining and Knowledge Discovery **1** (1):121-125
- Czyzewski, C. (1994). Artificial Intelligence-Based Processing of Old Audio Recordings. 97th Audio Engineering Society Convention, San Francisco, USA
- Fayyad, U., P. Smyth. (1993). Image Database Exploration: Progress and Challenges. Proceedings Knowledge Discovery in Databases Workshop. Washington, D.C., USA. pp14-27
- Fayyad, U., S.G. Djorgovski, N. Weir. (1996). Automating the Analysis and Cataloguing of Sky Surveys. In: U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy. (eds.). Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press. pp. 471-493
- Feldman, R., I. Dagan. (1995). Knowledge Discovery in Textual Databases (KDT). Proceedings 1st International Conference Knowledge Discovery and Data Mining. Montreal, Canada. pp112-117
- Han, J., M. Kamber. (2000). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers
- Matichuk, B., O.R. Zaïane. (2000). Unsupervised Classification of Sound for Multimedia Indexing. Proceedings 1st International Workshop on Multimedia Data Mining. pp31-36

- Noirhomme-Fraiture, M. (2000). Multimedia Support for Complex Multidimensional Data Mining. Proceedings 1st International Workshop on Multimedia Data Mining. pp54-59
- Papageorgiou, C., T. Evgeniou, T. Poggio. (1998). A Trainable Pedestrian Detection System. Intelligent Vehicles, Stuttgart, Germany
- Shearer, K., C. Dorai, S. Venkatesh. (2000). Incorporating Domain Knowledge with Video and Voice Data Analysis in News Broadcasts. Proceedings 1st International Workshop on Multimedia Data Mining. pp46-53
- Simoff, S.J., O.R. Zaïane. (2001). Report on MDM/KDD2000: The 1st International Workshop on Multimedia Data Mining. Submitted to SIGKDD Explorations, Newsletter of the Special Interest Group on Knowledge Discovery and Data Mining
- Tucakov, T., R. Ng. (1998). Identifying Unusual Spatio-Temporal Trajectories from Surveillance Videos. Proceedings of 1998 SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'98). Seattle, Washington
- Zaïane, O.R., J. Han, Z.N. Li, J. Hou. (1998a). Mining Multimedia Data. Proceedings CASCON'98: Meeting of Minds. Toronto, Canada. pp83-96
- Zaïane, O.R., J. Han, Z.N. Li, J. Chiang, S. Chee. (1998b). MultiMedia-Miner: A System Prototype for Multimedia Data Mining. Proceedings 1998 ACM-SIGMOD Conference on Management of Data. Seattle, Washington. pp581-583
- Zaïane, O.R., J. Han, H. Zhu. (2000a). Mining Recurrent Items in Multimedia with Progressive Resolution Refinement. Proceedings International Conference on Data Engineering (ICDE'2000a). San Diego, CA. pp461-470
- Zaïane, O.R., S.J. Simoff. (eds.). (2000b). Proceedings of the First International Workshop on Multimedia Data Mining (MDM/KDD2000), held in conjunction with ACM SIG-KDD conference. http://www.cs.ualberta.ca/~zaiane/mdm_kdd2000/
- Zaïane, O.R. (2000). Mining Multimedia Data. Invited Talk. XIV Brazilian Symposium on Databases (SBBD'2000). João Pessoa, Paraíba, Brazil

5.5.2 Musical audio mining

Marc Leman[1]

Introduction

Musical audio mining can be defined as data mining on musical audio. It will allow users to search and retrieve music, not only by means of text queries (such as title, composer, text song, conductor, orchestra), but also by means of content-based musical queries, such as query-by-humming/singing/playing, by specification of a list of musical variables (such as 'happy', 'energetic', etc.), or by means of given sound excerpts, or any combination of audio and text. Musical audio mining opens a number of perspectives for the music industry and related multimedia commercial activities. In this paper we introduce the reader to some basic concepts of musical audio mining and discuss its relevance for commerce and industry.

Musical audio mining in the information society

Music plays an important role in our high-tech information society. Music is nowadays present in every aspect of life, public and private, and thus has a very noticeable influence on people. It allows the expression of affect and emotion which is necessary for the formation of identity and the development of the well-being of the individual in a social environment. This involvement with music is reflected in an elaborate network of music commodities and services, called 'the music industry'. The core of this industry is dominated by a few major companies, followed by a large number of smaller companies [Pichevin, 1997].

Just to give a rough idea of the size of the market: in 1999, the global music retail market (music sales of singles, LPs, MCs, CDs) was worth about 38,506 million US$ with a total unit sales of 3,459 million units showing a growth by 2,6% a year in real value since 1991 [IFPI, 2000]. If, next to that, one considers the social and economical impact of the music recording sector, copyright associations, organizations for concerts and other live performances, musical instrument and consumer audio markets, music in broadcasting, and education, one quickly understands that the music 'business' relies on enormous human and financial resources. Data from Laing [1996] suggest that the music industry had a greater turnover than both the cinema and video industries, providing work in Europe alone to over 600,000 people, and having an impact on millions of people involved in education and business.

Despite its consumer market and large network of interconnected industrial activities, musical content has long been underestimated as a potential source of economical activity. But this is rapidly changing. Phenomena such as Napster

1 Prof Dr M. Leman, Marc.Leman@rug.ac.be, IPEM - Department of Musicology, Ghent University, Ghent, Belgium, http://www.ipem.rug.ac.be/

and Freenet have demonstrated the potential of distributed decentralized information storage and retrieval systems. Given the large quantity of mp3 audio files available, many people have become aware of the need for content-based search and retrieval.

The classical approach based on text as supplier of meta-data of musical audio, such as the name of the songwriter, the title and the text of the song, the musician, the director, the orchestra, recording date , etc. can be extended with descriptions based on the characteristics of audio. These descriptions can be generated partly manually, partly automatically. In combination with sophisticated profiling techniques, they offer the basis for future intelligent and flexible music mining systems. Given the nature of the problem, it goes without saying that the musical audio mining community is an interdisciplinary research community where musicologists, engineers, mathematicians, sociologists, and others have joined their forces [Leman, 2002].

Techniques for musical audio mining and retrieval

Music search and retrieval is often based not on audio, but on symbolic representations of music. A simple but powerful idea has been to conceive those symbolic representations in terms of the now familiar 'electronic scores'. MIDI (Music Instrument Digital Interface) is an example. This industrial standard for communication between electronic musical instruments and its encoding of music can be conceived of as a score, or note notation, in electronic format. But there are other formats for electronic scores as well, see [Selfridge-Field, 1997]. Many approaches in music information retrieval aim at reducing music first to a notated melody (as electronic score) and associate the proper audio files to this. Search is done on the melody, and retrieval can deliver the necessary audio files. It is often admitted that melodies may not be sufficient as an encoding of all kinds of music, but at least they provide an interesting paradigm for content-based music search and retrieval.

How, then are melodies related to musical content? A melody may be notated as a sequence of pitches or intervals (distances between pitches), and durations for each pitch (or distances between durations). That's fairly simple, and therefore appealing. Search in audio is then reduced to search on symbolic strings that handle pitch and duration. Yet, this simple encoding method already shows that musical content is not always easy to define.

Think of defining the notion of melodic similarity, which is a necessary component in any search and retrieval system. It is evident that proper accounts of musical perception, and perception-based modeling should be taken into account, when melodies are compared. The similarity of two strings is typically

calculated as the 'edit'-cost it takes to make the two strings identical by performing a deletion, insertion, and replacement operation on the characters. Characters, in this case, stand for musical notes, or durations, and sequences of characters define melodies. [Rolland, 1999] developed algorithms for discovering patterns in musical sequences, taking into account insertion, deletion and replacement of notes, as well as multiple notes. The operations are furthermore weighted depending on the level of abstraction that is used. This is indeed very typical for music: a melody is not just a concatenation of notes, but notes constitute relationships at different levels of abstraction. One may consider exact pitches, pitch intervals, or general melodic contour as three different abstraction levels related to pitch. But similar abstraction levels exist for rhythm (such as pulse, beat, meter, rhythm units). Furthermore, the meaning of a melody often depends on the tempo, some notes that are 'edited' are less important than others, depending on their position in the metric structure, and so on. In other words, similarity measurement is not just a matter of string comparison, but one may prefer to introduce high level musical knowledge to make comparisons relevant from the viewpoint of human perception. Musical content processing, therefore, is not that simple. Furthermore, systems may take into account user profiles [Rolland, 2001]. Music education or none? It may make a difference to how one deals with musical content.

Melody representations thus far have been very important, because they offer a major paradigm for music search and retrieval. The paradigm relies on both a music database and a query system. The retrieval will often pass via a symbolic encoding system, such as melody notation, although the representation is not necessarily limited to notated melodies. It can be based on different encoding techniques such as Hidden Markov Models (see Section 6.2.10) or neural networks (see Section 6.2.8). Nevertheless, the procedure is often similar: queries based on audio input (humming/sound examples) will first be transformed into a more abstract representation, making search and similarity measurement feasible. The data entries found can then be linked to the proper audio file which is retrieved.

In order to become acquainted with some of the techniques, we introduce two approaches based on electronic scores. In the next section, we discuss music mining based on audio files. A general observation is that many research groups focus on electronic scores, rather than on audio. This seems to indicate that audio research is in an initial phase, where basic concepts and techniques for audio need further development. Speech processing offers some inspiring paradigms, although one has to take into account that musical audio and speech are different in many aspects.

Music search and retrieval based on melodies

Some basic techniques of musical search and retrieval can be illustrated by means of practical applications.

The MELDEX[2] system is designed to retrieve melodies from a database on the basis of a few notes sung into a microphone, hummed, or otherwise entered. The acoustic input is transcribed into music notation and then a database is searched for melodies that contain similar patterns [Ghias, 1995; McNab, 1997]. The audio to melody transformation is based on the extraction of the fundamental pitch. Users are requested to sing with 'da' and 'ta' so that beginnings and endings of notes can be easily determined by segmentation on amplitude. Retrieving the music from the collection of musical scores is then essentially a matter of matching input strings against a database. To facilitate the task, users can constrain the search. For example, they can compare their input with the beginning or with the chorus of the music. Search is then based on interval directions (contour), rather than pitch ratios, or musical intervals, but the latter is an option. The technique comes down to the classical calculation of the 'edit'-cost by means of deletion, insertion, and replacement. The cost of a sequence of edit operations is the sum of the costs of the individual operations, and the aim is to find the lowest-cost sequence that accomplishes the desired transformation. Using dynamic programming, the optimal solution can be found in a time proportional to the product of the lengths of the sequences. As mentioned before, the search algorithms are improved, when knowledge of musical content processing can be included. Tunes with a certain percentage of matching are returned and the user can then download the files. Themefinder[3] is a similar search and retrieval system based on melodies, but it does not involve audio. Related work on melodies can be found in [Cambouropoulos, 2001].

An entirely different approach is based on so-called content-addressed memories. [Toiviainen, 2001][4] applies this idea to melodies, but it can be extended to any kind of symbolic encoding. The method draws on the findings that listeners judge the similarity of melodies on the basis of frequently occurring features of music. Common statistical measures of music, such as distributions of pitches, intervals and durations, pitch transitions, interval transitions, and duration transitions are extracted from each melody separately. Each melody is thus encoded by a set of variables defining a particular quality of the melodic content. Similarities among melodies can be studied on the basis of a comparison of their probability distributions. For example, when you are interested in a qualitatively high match of pitch sequences, you may focus on the variable (and its corresponding probability distribution) that is based on counting the occurrences of any 3 pitch successions. You may then proceed by performing a proper clustering of these distributions over all melodies. Similarity measurement will then typically be based on distance calculations in a multidimensional

2 http://www.nzdl.org/cgi-bin/library

3 http://www.themefinder.com

4 http://www.jyu.fi/musica/essen/

space. Constrained search can be realized by combinations of the appropriate distributions. Toiviainen and Eerola, for example, use the SOM [Self-Organizing Map, Kohonen, 1995] to cluster the melodies, and combinations of the variables of a melody were realized through the clustering of the position coordinates of the melody on different maps, each representing a single feature (see also Section 6.3.2, Self-Organizing Maps and [Kaski, 1997] on the CD-rom).

The statistical approach shows similarities with statistical techniques in speech processing [Jelenik, 1999] and is certainly worth further investigation. The technique is robust and could allow the representation of musical style.

Musical search and retrieval based on audio

In the previous section, we discussed approaches that reduce music to a notated melody. Of course, there is a price to be paid for this simplicity. First of all, not all music is notated by means of a score (think about ethnic music, electronic music, improvisations) and, secondly, the melody approach is not suited for polyphonic music or music with no explicit melodic character. As a consequence the technique is often restricted to the encoding of particular and well-documented folksongs and popular songs. It seems straightforward to expand the search techniques to polyphonic music notation, rather than melodic notation. This work is just starting, see [Dovey, 2001] who describes techniques for searching in polyphonic music notations, and [Doraisamy, 2001] who develops a search method based on all combinations of monophonic musical sequences. However, in dealing with much of the musical reality we need more advanced techniques to deal with audio.

Are there any techniques available which could handle any kind of musical audio? Transforming musical audio into some kind of symbolic or abstract representation is a necessary step for most musical search and retrieval systems. A straightforward paradigm for musical encoding is inspired by speech recognition research and runs as follows. The musical signal is first windowed into short (e.g. 30 ms) overlapping frames. For each of the frames, a short-time spectrum is computed and transformed into Mel Cepstrum Coefficients[5]. Alternatively, one may use an auditory model or wavelet transform. Then a Hidden Markov Model (HMM) can be used to classify the frames in an unsupervised way. The outcome will be an alphabet obtained by self-organization. In the study by [Aucouturier, 2001], the HMM learns different textures occurring in the music in terms of mixtures of Gaussian distributions[6] over the space of spectral envelopes. Learning is done with the Baum-Welsh algorithm, while the labeling is done using Viterbi decoding [Jelenik, 1999](see Section 6.2.10). Each state of the HMM then accounts for one texture. A particular musical piece will be encoded by means of state transitions, hence by a sequence of labels (corre-

5 A transformation with an algorithm which incorporates compensation for the variations in sensitivity of the human ear to different audible frequencies.

6 Indicates the probability density of values.

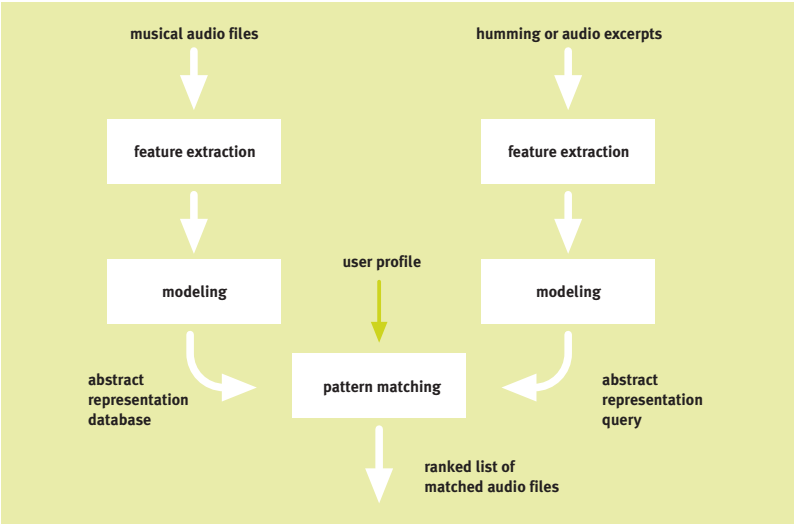


Figure 1
General diagram of musical audio mining based on audio queries.

sponding to the states), which we could call a 'state-score'. This is the classical technique borrowed from speech recognition (Figure 1).

Rather than comparing audio with audio directly, more abstract representations of musical audio are needed in order to take into account issues such as pitch transposition and tempo differences between the query and the stored music file. In the left part of the figure, musical audio files are transformed into abstract representations through feature extraction[7] and modeling. Feature extraction may produce pitch frames which are modeled, using HMM. The result is a kind of 'score' for each musical audio file with labels for pitches and some (flexible) representation for duration. The 'scores' make up a database which represents the musical audio files at a level of abstraction where pattern matching is possible. The query part has as input a musical audio file, which could be an excerpt of an existing piece of music, or a piece sung by the user in a microphone. A similar feature extraction and modeling can be done, which results in an abstract representation of the query. The pattern matching will be typically based on a similarity matching between the abstract representation of the query and the abstract representation of the music in the database. Pattern matching can be refined by a user profile. Many variations of this general schema are possible. Audio queries can be extended by a text-based music information retrieval system, different levels of search, and so on.

It goes without saying, however, that the definition of the acoustic events in music is not straightforward. Speech events are typically described in terms of a limited set of phonemes, and allow the statistical structures to be learned in a supervised way (i.e. on the basis of manual annotation). However, the acoustical events in music are much more numerous and should be learned in an unsu-

7 Extracting specific features from the original data.

pervised way, for example, by training HMM's on a large audio database. Once the HMM's have been derived, musical audio can be decomposed in terms of the states of the HMM. Instead of comparing raw audio material, similarity can then be based on the sequences of labels. In other words: appropriate string-matching can then be done by comparing 'state-scores'. Conceptually, they are on the same level as the notated melodies and therefore, similarity costs can be based on 'edit'-operations such as deletion, insertion, and replacement, using dynamic programming algorithms. Further research is needed, however, to demonstrate the feasibility of this paradigm.

Meanwhile, it is clear that variants of this schema can be explored. For example, one could first cluster the acoustical features using a neural network, and then represent the music as a trajectory in this clustering space. Trajectories can be learned and clustered so that similar music is projected at close distance in this space. Temporal characteristics can furthermore be dealt with by means of time-warping methods. [Mazzoni and Dannenberg, 2001], for example, extract a continuous pitch contour from a recording of the audio query and use then a time-warping algorithm to match this string of pitches against songs in a database of MIDI files.

Musart[8] [Birmingham, 2001] is an example of a query by a humming system based on the extraction of melodic contour and HMM's. Concerning sound effects search and query by example systems, see Musclefish[9]. [Blum, 1997] develops a variety of algorithms for analyzing the content of an audio signal and producing meta-data describing the signal. The first official beta release of the SoundFisher[10] program became available in August 2000. Follow-up versions are posted on a regular basis. The user can build complicated search requests based on any of the data, including the audio qualities or similarity to audio examples.

Quite a lot of other projects are related to query-by-humming. Just to mention a few: QBHAgent[11] is a system that employs the accepted notion of melodic pitch contours to support content-based navigation around a body of multimedia documents including MIDI and digital audio files. The system adopts an open hyper-media model, which enables the user to find available links from an arbitrary fragment of a piece of music, based on the content or location of that fragment. ECHO[12] is a software system that takes a sung melody as a search query. Bad singing could be adjusted for by calculating the difference in pitch and intervals between adjacent notes. Those calculations are then compared with music sequences in a database until a match is found. SoundCompass[13] is a query-by-humming system for a large music database. All processing of musical information is done on the basis of beats instead of notes. Tuneserver[14]

8 http://musen.engin.umich.edu

9 http://www.musclefish.com

10 http://www.soundfisher.com/

11 http://audio.ecs.soton.ac.uk/sbh

12 http://www.sonoda.net/

13 http://woodworm.cs.uml.edu/~rprice/ep/kosugi/

14 http://wwwipd.ira.uka.de/tuneserver/

attempts to match a wav-file containing a user's whistling to a classical music piece. Audio Logger[15] has an interactive query which interfaces with multiple modes such as a query by example (audio clip or music example), query-by-humming, query by concept, query by user drawn frequency and amplitude shape, query by instrument. Query by emotional conception of music: 'sorrow', 'sad', 'joy', query by pitch, query by frequency or amplitude shape, and query by keyword are also possible.

The University of Milan[16] has a project related to archiving, in collaboration with the Teatro alla Scala and the Bolshoi Theatre. Part of the project is devoted to the development of a content-based system [Haus, 2001]. Related groups in Europe are situated in different countries, among which IRCAM (France), and Pompeu Fabra University (Spain). CUIDADO[17] is an example of a EU-supported project, which aims at developing content-based technologies using MPEG 7 standard, building reusable modules for audio feature extraction, statistical indexing, database management, networking and constraint-based navigation. A project called MAMI (Musical Audio mining) has recently been established at Ghent University[18].

This list is not exhaustive and the reader may consult the recent proceedings of ISMIR[19] 2000 and 2001 for further details. In general, most audio-based systems so far deal with very limited databases. Large databases are typically encoded in melody notation.

MPEG-7 audio

Some words should be said about MPEG-7 audio. The background for this is that musical search and retrieval will be based on both automated and manual descriptions and that all this information needs to be represented in one way or another. Therefore, a recurrent idea has been to develop a standard for representing musical content, also called 'meta-data' in this context. This is the goal of MPEG-7 Audio, a multimedia content description interface, whose first version is available[20], but its stability has not yet been proven, and some of its representation types may raise questions from a musical point of view. Nevertheless, MPEG-7 Audio is an interesting initiative and a first attempt to develop a common representation for meta-data on music.
MPEG-7 Audio has been developing a description scheme for melodic information in terms of pitched, monophonic melodies. Another description scheme deals with musical instrument timbre, which is, according to this standard, either 'harmonic' or 'percussive'.
It is important, however, to understand that apart from some reference tools MPEG-7 does not come up with algorithms for, say, fundamental pitch extraction. Rather the aim is to provide a framework for the representation within a

15 http://www.inficad.com/~roz/audio.htm

16 http://woodworm.cs.uml.edu/~rprice/ep/haus/index.html

17 http://www.cuidado.mu/

18 http://www.ipem.rug.ac.be/mami/index.html

19 http://music-ir.org/

20 http://www.lis.e-technik.tu-muenchen.de/research/bv/topics/mmdb/e_mpeg7.html

larger context of feature representations and schemes. The MPEG home page[21] and the MPEG-7 Industry Focus Group web site[22] contain links to information about MPEG, links to other MPEG-7 web pages and many publicly available documents (the MPEG-7 standard is included on the CD-rom).

Understanding music for musical audio mining

Many audio-based approaches to music are characterized by a straightforward bottom-up approach and do not take music models into account in their processing. Most music researchers, however, argue that music needs to be described at different levels [Todd, 1991; Poli, 1991; Marsden, 1992; Balaban, 1992; Bigand, 1993; Haus, 1993; Roads, 1997; Leman, 1997; Zannos, 1998; Godoy, 2001], and that music like language obeys rules which are grounded in human biology and in cultural context. Progress in the acoustical analysis of music may be expected from the development of good music models, a better understanding of the musical structures underlying particular musical styles, and models of music perception. The following levels of description of music attempt to give an idea of the different possible approaches.

Signal representation

This is just how music is represented at the most basic level: as waveforms.

Frame-based representations

Frame based representations are derived from the waveform using shifted frames, as is nowadays done in many applications. Different types of frame-based representations can be related to different musical parameters such as energy, pitch, roughness, rhythm, spectrum, etc. The Mel Cepstrum Coefficients representation is popular for acoustical encoding, because it is a fast way to incorporate aspects of the human auditory system. In auditory modeling, which is often used as the acoustical front end in speech processing, the frame-based representations reflect neuronal activity at different time steps. [Leman, 2001] offers a toolbox for perception-based music analysis.

Parameter-based representations

These representations rely on modeling techniques over frame-based representations. An example is tonal tension, which can be calculated by comparing pitch images that reflect the development of musical pitch on a local scale on the one hand with pitch patterns of a global scale on the other hand [Leman, 2000]. Another straightforward technique is to use inter-onset-intervals as constrained time intervals from which parameters may be extracted. [Carreras, 1999] apply this technique to extract chord patterns from pitch images. In vibrato, the mean pitch pattern will be the pitch heard, but the pitch modulation may be as great as 60 cents (half a tone). [Rossigno, 1999] have worked out a feature

21 http://mpeg.telecomitalialab.com

22 http://mpeg-industry.com

extraction and temporal segmentation device. Characterizing the expressive content of musical signals is another 'hot' topic of research [Dannenberg, 1998]. The work of [Canazza, 1997] aims at extracting performance features from musical signals. These are just a few examples of a large range of features that can be extracted from musical signals.

Event and gesture-based representations

Events can be linked to the underlying parametric and frame-based representations. [Klapuri, 2001] has developed a system for the automatic transcription of music with the main emphasis on finding the multiple pitches of concurrent musical sounds. These techniques may be very useful for encoding, because they allow for the representation of musical content in terms of symbols: objects having attributes such as a particular beginning, end, duration, pitch, loudness, timbre, vibration frequency, vibration index, etc. Larger musical structures can be taken into consideration as well. An example is a repeating rhythm pattern, where the energy evolution over time represents the deployment of gestural characteristics in time. A musical content, therefore, may point to a melodic figure representing a particular musical thought or Gestalt. [Bresin, 2000] uses performance variables such as pitch number, inter-onset duration, off-time duration, sound level, pitch deviation, vibrato amplitude, vibrato frequency, etc. to define expressive performances in event-based music representations.

The concept-based representations

The next level then is purely conceptual. Concept-based representations provide descriptions for musical features in terms of the natural language such as 'rough', 'expressive', 'dissonant', 'graceful', etc. Primary concepts are those such as 'high', 'low', 'far', 'near', 'accellerando' or 'descellerando', which describe physical features of a musical pattern or event. At a second level, concepts may have a metaphorical origin rooted in space, movement, and effect. These are the so-called kinaesthetic [Laban, 1963], synaesthetic [Cytowic, 1989], and cenaesthetic concepts [Broeckx, 1981], which actually belong to another domain of experience, but which are often used to characterize music. Examples are 'press', 'glide', 'flick' or concepts such as 'labile', 'conflict', 'extinction'. Often these concepts can be connected to lower representational levels. Much more research is needed to understand how people use these concepts for retrieving music. Some related work has been done by [Pachet, 2001] who studied a method for parsing music file names.

To summarize: different representational levels for music exist, and they provide cues for meta-data representation of musical audio. What is needed is a conceptual architecture for music representation in which these different levels

can be handled in a flexible way. As far as I know, such a representational schema does not yet exist, though steps in the direction of multi-level architectures for musical representation are undertaken.

Economical and societal impacts

The economical and societal impacts of musical audio mining are primarily situated in the music industry, but they may prove to have an even wider impact as well. Results may be relevant to an entire domain of multimedia applications concerned with the automatic retrieval of information from databases incorporating files of different natures — such as text, speech, music and images — coexisting in the same software environment. In particular, the industries which have thus far been involved in text-based and speech-based data mining may learn how to extend their methodologies so as to include musical audio files in their applications as well.

In the present stage of development, musical audio mining research projects are upstream industrial research projects with a high potential in follow-up technological trajectories for diverse industrial players in the multimedia industry, in particular in the large and growing field of the data mining and content business (content creation, distribution, delivery, retrieval, consumption). Many of the tools envisioned in musical audio mining will prove valuable in the general domains of signal processing and pattern matching. In particular, we think of 'intelligent' interactive multimedia systems that work on the processing of musical content. Musical audio mining can further be used to retrieve songs from a database in a Karaoke machine, to shop for specific songs on CDs in a music store, or to request certain songs via a mobile phone. Research and industrial communities nowadays have a strong interest in non-verbal multimodal communication both with man and machine. The focus on non-verbal multimodal communication is of particular relevance to many areas in man-machine communication.

Related industry and commerce

Following is a more detailed view of the absorption of musical audio mining by the aimed sector in different application domains.

Music industry

The primary sectors of the music industry are: (a) the production and sale of sound recordings (b) the administration of authors rights and neighboring rights (c) the organization of concerts and other manifestations (d) the manufacture and sale of musical instruments. Musical audio mining research will have an impact on each of these:

– Musical audio mining is relevant for the production of sound recordings, because it has a focus on content. The current production techniques can be

improved so that more efficient content creation based on automated audio-analysis can be achieved. Search and retrieval of music on the basis of audio mining techniques is of course a core application.

- Techniques for feature extraction, statistical modeling, and similarity measurement can prove to be useful for companies such as SABAM, EMO and IFPI' which are involved in the protection of author's rights. Nobody questions the importance of an efficient copyright protection system, and good content extraction algorithms may be an indispensable means for the efficient tracing of eventual plagiarism.
- Concerts and performances can be advertised using content description techniques. The core technology of musical audio mining is furthermore highly relevant to the entertainment business. The (future) interactive discotheque is one example of a system in which groups of dancers interact and produce their music by dance in response to music machines. By dancing they can also influence the music which is produced by the system. Karaoke is another very popular field in which the musical audio mining techniques are central. In the Karaoke paradigm, a singer sings a popular song whose score is stored in the computer. The machine plays the accompaniment, and so as with the interactive discotheque, techniques of feature extraction and similarity measurement are of central importance in selecting the right music.
- As to the manufacture and sale of electronically enhanced music instruments, musical audio mining will be relevant in that it provides core technologies for the new generation of interactive multi-sensory audio/dance/video systems. The musical audio mining technology may serve to extract the appropriate content that is needed for man-machine interaction based on expressiveness [Camurri, 2001]. Applications are conceivable in the artistic cultural sector, for example in theater and opera houses, or in the commercial cultural sector such as discotheques, large multimedia manifestations, and very popular dance music environments. Applications which envision the interaction of large dancing masses with technology on the basis of content extraction in movement and audio, such as singing offers interesting opportunities.

Multimedia industry

- Apart from the specific musical applications mentioned in the previous section, there is a vast domain of possible applications where music is considered part of a larger information context.
- Musical audio mining is relevant for the search in multimedia archives, because audio is indeed an important aspect of many multimedia documents. Musical audio mining technology will provide the technology for separating for example songs from speech and from purely instrumental musical

passages. Access to sound sources of all kinds, including speech and music are very useful for researchers, for TV-production houses (e.g. finding appropriate musical sequences, or spoken or sung film/video fragments), as well as for individual consumers who want to get direct access to particular forms of music. Such systems will allow users to fast-forward through the news broadcast, stopping at all the occurrences of a given word or phrase, or avoiding musical intermezzos. By eliminating the need to listen to or transcribe lengthy recordings, this breakthrough capability can save enormous amounts of time and increase productivity. The capabilities of such systems can be greatly enhanced by more sophisticated tools for musical audio mining.

– Libraries, museums, foundations, universities make up some of the group of users that are focused on cultural heritage archives. Until now, rapidly growing multimedia databases offered only author/title/edition-like search and indexing capacities. This very general and crude method of archiving does not comply with modern needs of library users and modern technological possibilities. Musical audio mining technology will provide core technology for the development of new types of indexing and searching in multimedia archives, thus giving easier access to larger user groups. In connection to a film archive, for example, users may be interested in retrieving film fragments where similar musical sequences have been used.
– Other important multimedia users are recording studios, composers and new media artists, TV, radio, Internet-TV, radio and TV on-demand, and film production companies. The use of multi-purpose multimedia archives is a core aspect of their functioning. This category of users is dependent on easy access to custom multimedia databases, and their daily work produces hours of material, which has to be archived in a modern, efficient fashion. It is important to mention that the ever-growing role of Internet pushes technologies which will probably force a replacement of traditional TV and radio transmission with TV and radio on-demand. The success of these technologies will depend on the effectiveness of the embedded storage and retrieval systems. Again we claim that musical audio mining technology is a core technology in this field.

E-commerce applications

On-line shopping and E-commerce is an important new domain which will dominate the future retail market. As in the case of TV and radio-on-demand, musical audio mining offers core technologies to facilitate faster implementations of efficient and reliable archiving systems. Future promotion of record companies will depend on the accuracy with which the buyer can find the recording he is looking for. Present title/number catalogues for videos, for example, will be replaced by more human-friendly catalogues, which comply with ‘natural language’ search systems and ‘query-by-humming’ or ‘sounds like this song’.

Hardware manufacturers

If the musical audio mining research were to awaken a genuine need in the multimedia users, hardware manufacturers would follow with implementing compatible features on their hardware. In the audio/video home entertainment equipment field, one can imagine the video camera or DVD recorder automatically indexing recorded data with MPEG-7 descriptors. The same changes may be implemented in the fields of computer hardware, and electronic instruments.

Conclusion

Given the current state of the art in electronic-content delivery, the technological orientation of the music culture, and the interest of the music industry in providing musical commodities and services via the distributed electronic channels, there is an urgent need to develop advanced tools for music mining, that is, ways to deal with content concerning music and associated processing.

Musical audio mining is a young, but promising research field. However, due to the great variability of musical audio, its non-verbal basis, and its interconnected levels of description, musical audio mining is a rather difficult field and despite a growing group of researchers, it is still in its initial stage of development. Musical audio mining is rooted in musicology, where it draws on concept taxonomies that allow users to specify a musical piece in terms of more or less unique descriptors. These descriptors have their roots in acoustical properties of the musical audio, hence signal processing and statistical modeling are core disciplines as well because, they relate the audio to the conceptual taxonomy. As in text-based data mining, similarity measurement plays an important role in finding the appropriate connections between representational structures in the query and representational structures in the database.
The intrinsic interdisciplinarity as well as its foundation in musicology, sound and music computing, and similarity measurement give musical audio mining the status of a core research domain in multimedia. Given its relevance for the music industry and related commercial applications in multimedia, audio mining can be considered a very promising research field.

Acknowledgements

Thanks to M. Lesaffre and F. Carreras for their useful comments.

References

– Aucouturier, J., M. Sandler. (2001). Using Long-Term Structure to Retrieve Music: Representation and Matching. Proceedings of the 2nd Annual International Symposium on Music Information Retrieval. Indiana University Press, Bloomington, USA
– Balaban, M., K. Ebcioglu, O. Laske. (eds.). (1992). Understanding Music with

AI: Perspectives on Music Cognition. The MIT Press, Cambridge, MA, USA
- Bigand, E., S. McAdams. (1993). Thinking in Sound: the Cognitive Psychology of Human Audition. Oxford University Press, Oxford, UK
- Birmingham, W., R. Dannenberg, G. Wakefield, M. Bartsch, D. Bykowski, D. Mazzoni, C. Meek, M. Mellody, W. Rand. (2001). Musart: Music Retrieval via Aural Queries. Proceedings of the 2nd Annual International Symposium on Music Information Retrieval. Indiana University Press, Bloomington, USA, pp73-81
- Blum, T., D. Keisler, J. Wheaton, E. Wold. (1997). Audio Databases with Content-Based Retrieval. In: M. Maybury (ed.). Intelligent Multimedia Information Retrieval. MIT/AAAI Press, Cambridge, MA, USA
- Bresin, R., A. Friberg. (2000). Emotional Coloring of Computer Controlled Music Performance. Computer Music Journal. Computer Music Journal **24** (4):44-63
- Broeckx, J. (1981). Muziek, ratio en affect – over de wisselwerking van rationeel denken en affectief beleven bij voortbrengst en ontvangst van muziek. Metropolis, Antwerp, Belgium
- Cambouropoulos, E. (2001). Melodic Cue Abstraction, Similarity, and Category Formation: A Formal Model. Music Perception **18** (3):347-370
- Camurri, A., G. De Poli, M. Leman. (2001). A Multi-Layered Conceptual Framework for Expressive Gesture Applications. Proceedings of the Workshop on Current Research Directions in Computer Music. Audiovisual Institute, Pompea Fabra University, Barcelona, Spain
- Canazza, S., G. De Poli, S. Rinaldin, A. Vidolin. (1997). Sonological Analysis of Clarinet Expressivity. In: M. Leman. (ed.). (1997). Music, Gestalt, and Computing — Studies in Cognitive and Systematic Musicology. Springer Verlag, Berlin, Germany. pp431-440
- Carreras, F., M. Leman, M. Lesaffre. (1999). Automatic Description of Musical Sgnals Using Schema-Based Chord Decomposition. Journal of New Music Research **28** (4):310-331
- Cytowic, R. (1989). Synesthesia. Springer Verlag, Berlin, Germany
- Dannenberg, R., G. De Poli. (eds.). (1998). Synthesis of Performance Nuance. Special Issue of Journal of New Music Research, Swets & Zeitlinger, Lisse, The Netherlands
- Doraisammy, S., S. Rüger. (2001). An Approach towards a Polyphonic Music Retrieval System. Proceedings of the 2nd Annual International Symposium on Music Information Retrieval. Indiana University Press, Bloomington, USA. pp187-193
- Dovey, M. (2001), A Technique for 'Regular Expression' Style Searching in Polyphonic Music. Proceedings of the 2nd Annual International Symposium on Music Information Retrieval. Indiana University Press, Bloomington, USA. pp179-185

- Ghias, A., J. Logan, D. Chamberlin, B. Smith. (1995). Query-by-Humming – Large Musical Information Retrieval in an Audio Database. In Proceedings of the Third ACM International Conference on Multimedia. ACM-95, San Francisco, California, USA
- Godoy, R., H. Jorgensen. (eds.) (2001). Musical Imagery. Swets & Zeitlinger, Lisse, The Netherlands
- Haus, G. (ed.). (1993). Music Processing. Oxford University Press, Madison, A-R Editions, Wisconsin, USA
- Haus, G., E. Pollastri. (2001). An Audio Front End for Query-by-Humming Systems. Proceedings of the 2nd Annual International Symposium on Music Information Retrieval. Indiana University Press, Bloomington, USA. pp65-72
- IFPI. (2000). The Recording Industry in Numbers 2000. IFPI, London
- Jelenik, F. (1999). Statistical Methods for Speech Recognition. The MIT Press, Cambridge, MA, USA
- Klapuri, A. (2001). Automatic Transcription of Music. Proceedings of the Fourteenth Meeting of the FWO Research Society on Foundations of Music Research – Stochastic Modeling of Music. IPEM, Ghent University, Ghent, Belgium
- Kohonen, T. (1995). Self-organizing Maps. Springer Verlag, Heidelberg
- Laban, R. (1963). Modern Educational Dance. MacDonald & Evans, London, UK
- Laing, D. (1996). The Economic Importance of Music in the European Union. (http://www.icce.rug.nl/~soundscapes/DATABASES/MIE/Part1_introduction.html)
- Leman, M. (ed.). (1997). Music, Gestalt, and Computing – Studies in Cognitive and Systematic Musicology. Springer Verlag, Berlin, Germany
- Leman, M. (2000). An Auditory Model of the Role of Short-Term Memory in Probe-Tone Ratings. Music Perception **17** (4):481-509
- Leman, M., M. Lesaffre, K. Tanghe. (2001). A Toolbox for Perception-Based Music Analysis. Ghent: IPEM - Department of Musicology, Ghent University, Ghent, Belgium. http://www.ipem.rug.ac.be/Toolbox/index.html
- Leman, M. (2002). Systematic Musicology in the Information Society – Tendencies, Perspectives and Opportunities for Musical Content Processing. Manuscript in Press
- Marsden, A., A. Pople. (eds.). (1992). Representations and Models in Music. Academic Press, London, UK
- Mazzoni, D., B. Dannenberg. (2001). Melody Matching Directly from Audio. Proceedings of the 2nd Annual International Symposium on Music Information Retrieval. Indiana University Press, Bloomington, USA. pp17-18
- McNab, R.J., L.A. Smith, D. Bainbridge, I.H. Witten. (1997). The New Zealand Digital Library. D-Lib Magazine. May
- Pachet, F., D. Laigre. (2001). A Naturalist Approach to Music File Name

Analysis. Proceedings of the 2nd Annual International Symposium on Music Information Retrieval. Indiana University Press, Bloomington, USA. pp51-58
- Pichevin, A. (1997). Le Disque à l'Heure d'Internet: l'Industrie de la Musique et les Nouvelles Technologies de Diffusion. L'Harmattan, Paris, France
- Poli, G. de, A. Piccialli, C. Roads. (eds.). (1991). Representations of Musical Signals. The MIT Press, Cambridge, MA, USA
- Roads, C., G. de Poli, S. Pope. (eds.). (1997). Musical Signal Processing. Swets & Zeitlinger, Lisse, The Netherlands
- Rolland, P. (1999). Discovering Patterns in Musical Sequences. Journal of New Music Research **28** (4):334-350
- Rolland, P. (2001). Adaptive User Modeling in a Content-Based Music Retrieval System. Proceedings of the 2nd Annual International Symposium on Music Information Retrieval. Indiana University Press, Bloomington, USA. pp27-30
- Rossignol, S., X. Rodet, J. Soumagne, J.-L. Collette, P. Depalle. (1999). Automatic Characterization of Musical Signals: Feature Extraction and Temporal Segmentation. Journal of New Music Research **28** (4):281-295
- Selfridge-Field, E. (ed.). (1997). Beyond MIDI. The Handbook of Musical Codes. The MIT Press, Cambridge, MA, USA
- Todd, P., D. Loy. (eds.). (1991). Music and Connectionism. The MIT Press, Cambridge, MA, USA
- Toiviainen, P., T. Eerola. (2001). A Method for Comparative Analysis of Folk Music Based on Musical Feature Extraction and Neural Networks. Proceedings of the International Symposium on Systematic and Comparative Musicology III International Conference on Cognitive Musicology 2001. Department of Musicology, Jyväskylä, Finland
- Zannos, I. (ed.). (1998). Music and Signs. ASKO Art & Science, Bratislava, Slovakia

5.5.3 Image mining

Theo Gevers[1]

Introduction

In 1751-1772 Diderot and d'Alembert wrote the first full-size encyclopedia. When writing this encyclopedia, one of the major problems was that many real-world concepts and entities such as musical instruments, tools and pieces of furniture were difficult or impossible to describe in words alone. For this reason, they included a large number of pictures (i.e. sketches and drawings) of the real-world concepts in their edition. The idea of using pictures for object description and archiving was continued in other encyclopedias. In fact, the encyclopedia of Diderot and d'Alembert can be seen as one of the first attempts at using picture information for object description and archiving.

Today, very large digital image archives have been created and used in a number of applications including archives of images of postal stamps, textile patterns, museum objects, trademarks and logos, and views from everyday life as it appears in home videos and consumer photography. Moreover, with the growth and popularity of the World Wide Web, a tremendous amount of visual information has been made publicly accessible. As a consequence, there is a growing demand for search and mining methods to access pictorial entities in large image archives.

Currently, a large number of text-based search engines are available and they have been proven to be very successful in searching documents. To locate pictorial information, these text-based search engines assume that textual descriptions of the visual data are present. However, people are reluctant to categorize visual information verbally. This argument holds especially for images available on the Internet. Moreover, using text as the basis for image retrieval is almost always inadequate due to the semantic richness of pictorial information as was already recognized by Diderot and d'Alembert. Hence, the capabilities of current text-based search engines for retrieving images are limited.

To this end, content-based image retrieval systems have been developed based on multiple features (e.g. color, shape and texture) describing the image content [Visual, 1999; CPVR, 2001; Flickner, 1995; Pentland, 1994; Smeulders, 2000], for example. In fact, these systems retrieve images entirely on the basis of pictorial information. Most of these content-based search systems use the so-called query by example paradigm. The basic idea of image retrieval through query by example is to extract low-level, salient features from images in the database which are stored and indexed. This is done off-line. These features are typically derived from shape, texture or color information, and will be discussed later. The on-line image retrieval process consists of a query example image, given by the user on input, from which image features are extracted. These

1 Dr T. Gevers, gevers@science.uva.nl, ISIS, Faculty of Science, The Universiteit van Amsterdam, Amsterdam, The Netherlands, http://carol.wins.uva.nl/~gevers

image features are used to find images in the database which are most similar to the query image. A drawback, however, is that these low-level image features are often too restricted to describe images on a conceptual or semantic level. This semantic gap is a well-known problem in content-based image retrieval (see also Inset 1).

Inset 1: Semantic gap

[Hanjalic, 2001] explains the semantic gap in a wider context accordingly: typical retrieval tasks, such as 'find me an image (or a video clip) showing a bird!', are formulated at a cognitive level, according to the human capability of understanding the image or video content and analyzing it in terms of semantic elements such as the meaning (role) of objects, persons, sceneries, thematic (story) segments, or the context of an image or video segment.

In contrast, the information that can be extracted from an image or a video at algorithmic or system level is rather technical than cognitive and consists of low-level features.

For general-purpose image retrieval, query formulation is often not easy to understand for users without any image processing background. In particular, most users find it difficult to formulate queries that are well designed for image retrieval purposes: 'How do I formulate a query to retrieve an image of a bear.' Due to inadequate query formulation, the system will retrieve (too) many non-relevant images. Further, image-processing background is required when interpreting the retrieval results: 'Why did the system retrieve a logo-image of a red car when I searched for a logo-image containing a bear.' In this case, a system should make clear the reason (how and why) the images have been retrieved and consequently to enable the user to adjust/improve the pictorial query in an intuitive and interactive manner. Therefore, recent research is mainly focused on user interaction to select relevant candidate images to refine the query. Further, user friendly visualization tools are being researched indicating why and how the query formulation has matched the retrieved image.

Image classification

To enhance the performance of content-based retrieval systems (e.g. by pruning the number of candidate images), image classification could be applied, prior to the image retrieval process, to group images into semantically meaningful classes [Vailaya, 1999; Vailaya, 1999; Yu, 1995; Zhong, 1995]. The advantage of these image classification schemes is that simple, low-level image features can be used to express semantically meaningful classes. Image classification could be based on unsupervised learning techniques such as Clustering (see Section 6.2.6), Self-Organizing Maps (SOM, See Section 6.3.2) [Zhong, 1995] and Hidden Markov models (See Section 6.2.10) [Yu, 1995]. Further, supervised

grouping can be applied. For example, vacation images have been classified based on a Bayesian framework into city versus landscape by supervised learning techniques (See Section 6.2.11) [Vailaya, 1999]. However, these classification schemes are entirely based on pictorial information. Aside from image retrieval [Favella, 1999; Sclaroff, 1999], very little attention has been paid on using both textual and pictorial information for classifying images on the Web. This is even more surprisingly, if one realizes that images on web pages are usually surrounded by text and discriminatory HTML tags such as IMG, and the HTML fields SRC and ALT. Hence, web images have intrinsic annotation information induced by the HTML structure. New research is directed toward the use of both textual and pictorial information for classifying images.

Interactive image retrieval by relevance feedback

Due to the semantic gap, most of the content-based retrieval systems today incorporate some kind of relevance feedback. Relevance feedback is an automatic process designed to produce improved query formulations following an initial retrieval operation. The effect of such query alteration process is to 'move' the query in the direction of the relevant images and away from the non-relevant ones. An overview of a standard retrieval system with relevance feedback mechanism is given in Figure 1.

The major components of the system are described below.

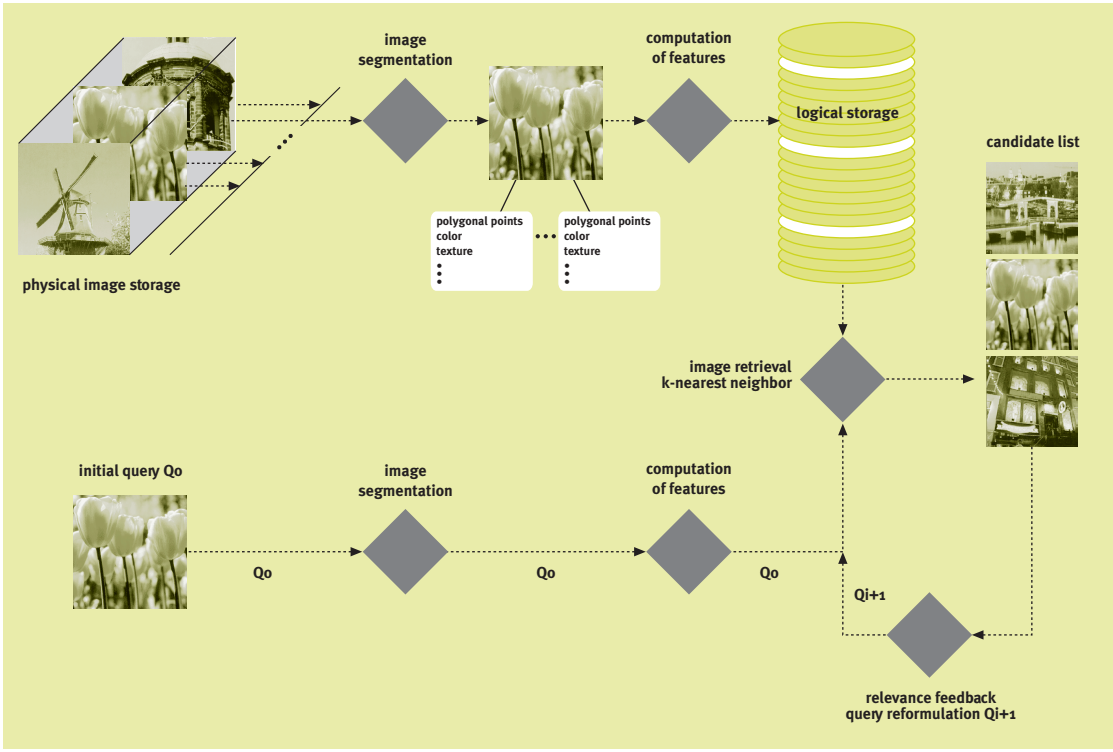


Figure 1
Overview of a content-based image retrieval system using relevance feedback.

Features are extracted automatically from the images in the database which are stored and indexed. This is done off-line. The on-line image retrieval process consists of a query example image from which image features are extracted. These image features are used to find the images in the database which are most similar. Then, a candidate list of most similar images is shown to the user. From the user-feedback the query is optimized and used as a new query.

Interactive query formulation

Interactive query formulation is offered either by query (sub)image(s) or by offering a pattern of feature values and weights. To achieve interactive query formulation, an image is sketched, recorded or selected from a repository. This is the query definition with the aim of finding a similar image in the database. Note that 'similar image' may imply a partially identical image, or a partially identical object in the image.

Image features

Image feature extraction is an important step in image indexing and search. In fact, a concise and complete set of image features should be provided. Further, image features should have high discriminative power.

Color

Various color-based image search schemes have been proposed based on different representation schemes such as color histograms, color moments, color edge orientation, color texture, and color correlograms [Visual, 1999; CVPR, 2001]. These image representation schemes have been created on the basis of RGB, and other color systems such as HSI and CIE L*a*b*. In particular, the Picasso [Bimbo, 1998] and ImageRover [Sclaroff, 1997] system use the CIE L*u*v* color space for image indexing and retrieval. The QBIC system [Flickner, 1995] evaluates similarity of global color properties using histograms based on a linear combination of the RGB color space. MARS [Servetto, 1998] is based on the CIE L*a*b* color space which is (like CIE L*u*v*) a perceptual uniform color space. The PicToSeek system [2] [Gevers, 2000] is based on color models robust to a change in viewing direction, object geometry and illumination.
From the systems above, it can be seen that the choice of color systems is of great importance to proper image retrieval. However, no color system can be considered as universal, because color can be interpreted and modeled in different ways. Each color system has its own set of color models, which are the parameters of the color system. Color systems have been developed for different purposes: 1. Display and printing processes: RGB, CMY. 2. Television and video transmission efficiency: YIQ, YUV. 3. Color standardization: XYZ. 4. Color uncorrelation: I1I2I3. 5. Color normalization and representation: rgb, xyz. 6. Perceptual uniformity: U*V*W*, L*a*b*, L*u*v*. With this large variety of color

2 http://carol.wins.uva.nl/~gevers/PicToSeek/

systems, the inevitable question arises which color system to use for which kind of image retrieval application. To this end, a survey is given by [Gevers, 2001] on the basics of color, color models and ordering systems, and the state of the art on color invariance. Further, a color system taxonomy is provided which can be used to select the proper color system for a specific application.

Texture

Texture can be described by its color primitives and their spatial layout. The spatial layout can be periodic, quasi-periodic or random. Thus, texture operators can be divided into two groups: structural and statistical. Surveys on texture are given by [Haralick, 1979] and [Gool, 1985]. If the goal is to retrieve images containing objects having irregular texture organization, the spatial organization of these texture primitives are, in the worst case, random. In this case, it is better to focus on statistical texture measures. It has been demonstrated that for irregular textures, the comparison of gradient distributions achieve satisfactory accuracy [Gorkani, 1994; Ojala, 1996; Pietikainen, 1996] as opposed to fractal or wavelet features.

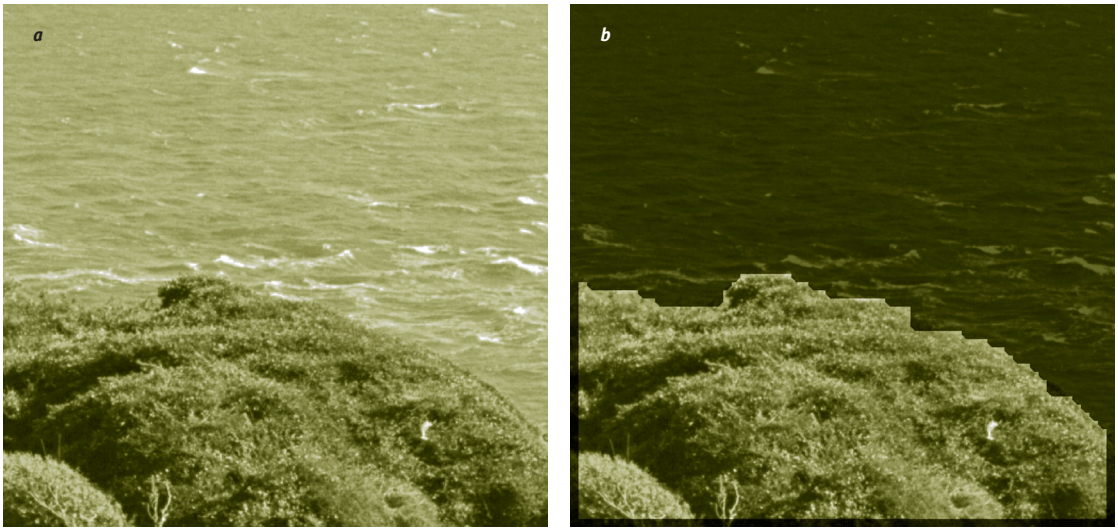


Figure 2
Texture examples.
a *An image showing irregular textured regions (e.g water and grass).*
b *Segmentation of the textured grass region based on gradient distributions.*

Shape

Because the physical world is three-dimensional and an image contains a two-dimensional projection of this reality, recognizing 3-D objects from 2-D projective images is a difficult problem, because the projection loses some 3-D information. In order to succeed, two major interrelated problems should be addressed: object representation and matching. The representation should be good enough to allow reliable and efficient matching. The recognition consists of matching the stored models, encapsulated in a representation scheme, against the image to determine which model corresponds to which portion of an

image. Therefore, the object models should be acquired automatically or manually. One way to proceed is to approximate these object models by polygons. The polygons are then represented by a sequence of projective invariant shape descriptors where the dimensional number of the sequence is the vertex number of the planar polygon. The sequence of shape descriptors of each planar polygon is then coded and entered into a hash table[3] yielding a mechanism for efficient indexing and retrieval. Once the database of object models are encoded and stored into the hash table which is done off-line, the on-line recognition consists of segmenting the image into polygon approximations, encoding them and use them to check the appropriate entry in the hash table, and for every model that appears a vote is given for the model corresponding to the one in the image. Hypotheses are obtained for a particular model, when the number of votes exceeds a certain threshold. Finally, the verification step consists of applying a least square match between the model and the instance.

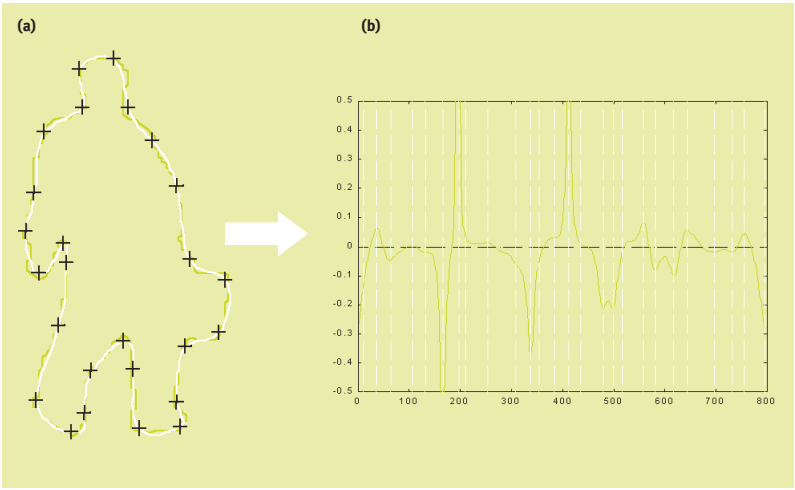


Figure 3
Example of shape representation. Object shape (a) and the corresponding curvature function (b) (Source: A. Hanjalic).

Representation

In general, the image feature sets are represented by an n-dimensional feature space. In this way, the domain dependent part of the whole image retrieval system is reduced to a minimum. Weights can be assigned corresponding to the feature frequency giving the well-known histogram form, where the feature frequency is the frequency of occurrences of a specific image feature value in the image or query (e.g. the total number of red pixels in an image). However, for accurate image object search, it is desirable to assign weights in accordance to the importance of the image features. For example, the image feature weights used for both images and queries can be computed as the product of the features frequency multiplied by the inverse collection frequency factor. In this way, features are emphasized having high feature frequencies, but low overall collection frequencies.

3 Table consisting of number representations of features, indexed by these numbers (see Section 6.2.12)

Matching measures

The actual matching process can be seen as a search for the k elements in the stored image set closest to the query image. As both the query images as the data set is captured in feature values, the similarity function operates between the weighted feature sets. Again, to make the query effective, attention has to be paid to the selection of the similarity function. A proper similarity function should be robust to object fragmentation, occlusion and clutter by the presence of other objects in the view. For example, it is known that the mean square and the Euclidean similarity measure provide accurate retrieval without any object clutter[4] [Gevers, 2000; Swain, 1991]. This similarity measure has a maximum of unity that occurs if and only if the query exactly matches the image in the database. Accurate image retrieval, when using this similarity function, is obtained as a result of the fact that this similarity function is symmetric and can be interpreted as the number of pixels with the same values in the query image which can be found present in the retrieved image and vice versa. However, the Euclidean similarity measure provides very poor retrieval results in the context of object occlusion and clutter. In this case, histogram intersection is more suited.

Search strategies

In the field of pattern recognition, several methods have been proposed that improve classification automatically through experience such as artificial neural networks (Section 6.2.8), decision tree learning (Section 6.2.7), Bayesian learning (6.2.11) and k-nearest neighbor classifiers. Except for the k-nearest neighbor classifier, the other methods construct a general, explicit description of the target function, when training examples are provided. In contrast, k-nearest neighbor classification consists of finding relationship to the previously stored images each time a new query image is given. When the user gives a new query, a set of similar related images is retrieved from the image database and used to classify the new query image. The advantage of k-nearest neighbor classification is that the technique constructs a local approximation to the target function that applies in the neighborhood of the new image query images, and never constructs an approximation designed to perform well over the entire instance space.

Because the k-nearest neighbor algorithm delays classification until a new query is received, significant computation can be required to process each new query. Various methods have been developed for indexing the stored images so that the nearest neighbors can be identified efficiently at some additional costs in memory, such as a k-d tree, R*-tree or a SS-tree, [Guttman, 1984] for example. Unfortunately, the complexity of these search algorithms grows exponentially with the dimension of the vector space, making them impractical for dimensionality above 15. Such high dimensionality can be expected in practice for general-

4 The presence of other objects in the view surrounding the object to be retrieved.

purpose image retrieval. Therefore, new indexing and data structures are being researched to improve performance in high dimensional spaces.

Visualization

Visualization of the feature matching results gives the user insight in the importance of the different features (even for those not used by the user during the previous search). To achieve this, the local back projection algorithm can be used on the basis of histogram intersection [Swain, 1991] for example.

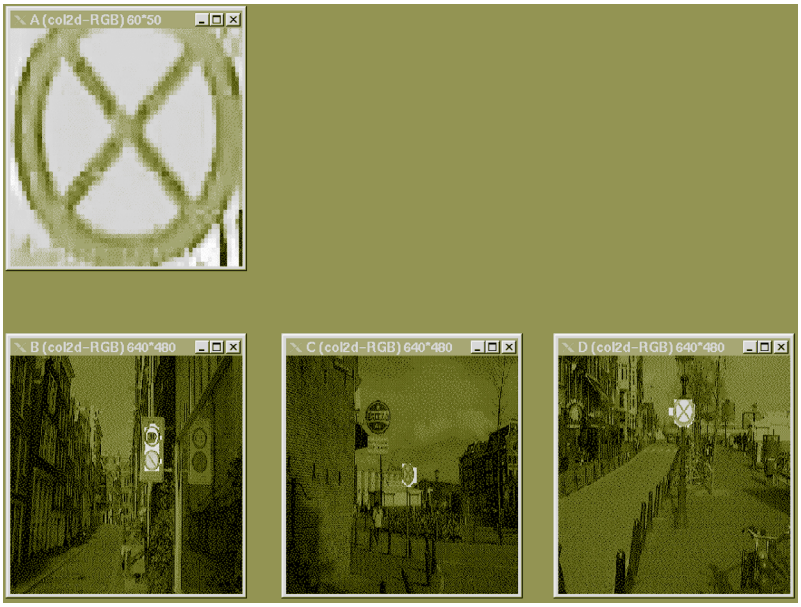


Figure 4
Visualization of the feature matching process. A traffic sign and corresponding colors found in the images similar to that of the traffic sign.

Windowing and information display techniques can be used to establish communications between system and user. In particular, screen pointers are used to designate certain images as relevant to the user's needs. These relevance indications are then further used by the system to construct modified feedback queries.

Relevance feedback

Relevance feedback is an automatic process designed to produce improved query formulations following an initial retrieval operation. Relevance feedback is needed for image retrieval where users find it difficult to formulate pictorial queries, which are well designed for accurate retrieval purposes. For example, without any specific query image example, the user might find it difficult to formulate a query (e.g. to retrieve an image of a car) by an image sketch or by offering a pattern of feature values and weights. This suggests that the first search operation should be conducted with a tentative, initial query formulation, and should be processed as a trial search only with the aim of retrieving only a few

useful images from the large image collection. These initially retrieved images should then be examined for relevance, and a (new) improved query formulation should be constructed with the purpose to retrieve more relevant images in subsequent search operations. Hence, from the user feedback giving negative or positive answers, the method can automatically learn which image features are more important. The system uses the feature weighting given by the user to find the images in the image database which is most similar with respect to the feature weighting. The feedback process can be represented graphically as a migration of the query vector from one area to another in the n-dimensional space.

Database technology

Research in database management has reached a state where relational database systems are readily available to manage large amounts of data. The database community estimates that at least 80% of all data still resides outside the confines of a database management system, i.e. in bulk stores (audio, video, images) and files (word processing, consumer use). Therefore, the common opinion is that in the future research should be more focused on bridging the gap between image processing and database technology.

Testing the system[5]

In general, image search systems are assessed in terms of precision, recall, query-processing time as well as reliability of a negative answer. Further, the relevance feedback method is assessed in terms of the number of iterations to approach to the ground-truth (the set of objects definitely satisfying the search). Today, more and more images are archived yielding a very large range of complex pictorial information. In fact, the average number of images, used for experimentation as reported in the literature, increased from a few in 1995 to over a million in 2002. It is important to note that the dataset should have ground-truths i.e. images which are annotated to be (non)relevant to a given query. For example, if one is looking for a specific object in an image, then the relevant images should contain at least part of the object to be searched for. In general, it is hard to generate these ground-truths, especially for very large datasets.

Conclusion and future trends

Content-based image retrieval and classification has reached a mature state and various commercial products have been put into the market [QBIC-IBM, 1995][6], for example.

From a scientific perspective the following trends can be distinguished. First, large scale image databases are being created. Obviously, large scale datasets provide different image mining problems to rather small, narrow-domain

5 see http://www.benchathlon.net/

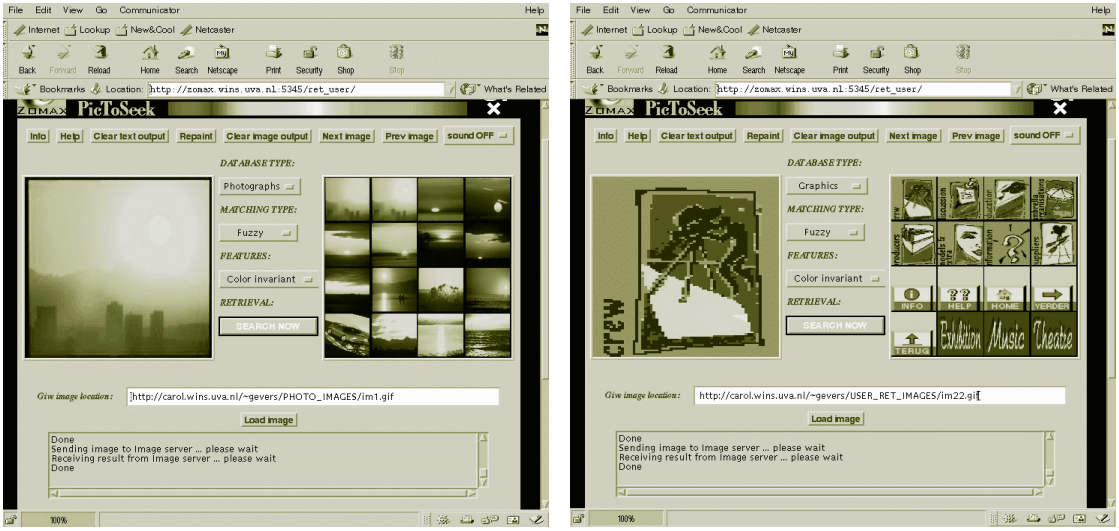
6 http://wwwqbic.almaden.ibm.com/

datasets. Second, research is directed towards the integration of different information modalities such as text, pictorial, and motion. Third, relevance feedback will be and still is an important issue. Finally, invariance is necessary to get to general-purpose image retrieval.

From a societal/commercial perspective, it is obvious that there will be an enormous increase in the amount of digital images used in various communication frameworks such as promotion, sports, education, and publishing. Further, digital images have become one of the major multimedia information sources on Internet, where the amount of video on the Web is growing by 350,000 hours per week today. Moreover, with the introduction of the new generation UMTS cell-phones, a tremendous market will be opened for the storage and management of pictorial data. Due to this, tremendous amounts of pictorial information, image mining and search tools are required, as indexing, searching and assessing the content of large scale image databases is inherently a time-consuming operation, when done by human operators. Therefore, product suites for content-based video indexing and searching are not only necessary, but essential for future content owners in the field of entertainment, news, education, video communication and distribution.

I am sure that you get the picture...

Figure 5
Illustration of the PicToSeek system where the typical application is considered of retrieving images containing an instance of a given object or genre. PicToSeek: http://carol.wins.uva.nl/~gevers/PicToSeek/.



References

- Bimbo, A. del, M. Mugnaini, P. Pala, F. Turco (1998). Visual Querying by Color Perceptive Regions. Pattern Recognition **31** (9):1241-1253
- CVPR2001. (2001). Proceedings of IEEE Workshop on Content-Based Access and Video Libraries. CVPR, Hawaii
- Favella J., V. Meza. (1999). Image-Retrieval Agent: Integrating Image Content

and Text. IEEE Intelligent Systems
- Flickner, M. et al. (1995). Query by Image and Video Content: the QBIC System, IEEE Computer **28** (9)
- Gevers, Th., A.W.M. Smeulders. (1999). Color Based Object Recognition. Pattern Recognition **32**. pp453-464
- Gevers, Th., A.W.M. Smeulders. (1999). Content-based Image Retrieval by Viewpoint-Invariant Image Indexing. Image and Vision Computing **17** (7)
- Gevers, Th., A.W.M. Smeulders. (2000). PicToSeek: Combining Color and Shape Invariant Features for Image Retrieval. IEEE Trans. on Image Processing **9** (1):102-120
- Gevers, Th. (2001). Color Based Image Retrieval. In: M. Lew (ed.). Principles of Visual Information Retrieval. Springer Verlag
- Gool, L. van, P. Dewaele, A. Oosterlinck. (1985). SURVEY: Texture Analysis anno 1983. CVGIP **28**:336-357
- Gorkani, M., R. Picard. (1994). Texture Orientation for Sorting Photos at a Glance. Proceedings IEEE Conference on PR. pp459-464
- Guttman, A. (1984). R-trees: A Dynamic Index Structure for Spatial Searching. ACM SIGMOD. pp47-57
- Haralick, R.M. (1979). Statistical and Structural Approaches to Texture. Proceedings IEEE 67:786-804
- Ojala, T., M. Pietikainen, D. Harwood. (1996). A Comparison Study of Texture Measures with Classification based on Feature Distributions. Pattern Recognition **29**:51-59
- Pietikainen, M., S. Nieminen, E. Marszalec, T. Ojala. (1996). Accurate Color Discrimination with Classification based on Feature Distributions. Proceedings ICPR **3**:833-838
- Pentland, A., R.W. Picard, S. Sclaroff. (1994). Photobook: Tools for Content-based Manipulation of Image Databases. Proceedings of Storage and Retrieval for Image and Video Databases II **2** (185). SPIE, Bellingham, Washington. pp34-47
- Sclaroff, S., L. Taycher, M. la Cascia. (1997). ImageRover: A Content-based Image Browser for the World Wide Web. Proceedings of IEEE Workshop on Content-based Access and Video Libraries
- Sclaroff, S., M. la Cascia, S. Sethi, L.Taycher. (1999). Unifying Textual and Visual Cues for Content-based Image Retrieval on the World Wide Web. CVIU **75** (1/2):86-98
- Servetto, S., Y. Rui, K. Ramchandran, T.S. Huang. (1998). A Region-Based Representation of Images in MARS. Journal on VLSI Signal Processing Systems **20** (2):137-150
- Smeulders, A.W.M., M. Worring, S. Santini, A. Gupta, R. Jain. (2000). Content-based Image Retrieval at the End of the Early Years. IEEE Transactions on Pattern Analysis and Machine Intelligence

– Swain, M.J., D.H. Ballard. (1991). Color Indexing. International Journal of Computer Vision **7** (1):11-32
– Vailaya, A., M. Figueiredo, A. Jain, H. Zhang. (1999). Content-Based Hierarchical Classification of Vacation Images. IEEE International Conference on Multimedia Computing and Systems. June 7-11
– Vailaya, A., A. Jain, H. Zhang. (1999). A Bayesian Framework for Semantic Classification of Outdoor Vacation Images. Proceedings of IS&T/SPIE Storage and Retrieval for Image and Video Databases VII 3656
– Visual99. (1999). Proceedings of Visual99 Information Systems: The Third International Conference on Visual Information Systems, Amsterdam, The Netherlands
– Yu, H.-H., W. Wolf. (1995). Scene Classification Methods for Image and Video Databases. Proceedings SPIE on Digital Image Storage and Archiving Systems, San Jose, CA. pp363-371
– Zhong, D., H.J. Zhang, S.-F. Chang. (1995). Clustering Methods for Video Browsing and Annotation. Proceedings SPIE on Storage and Retrieval for Image and Video Databases, San Jose, CA

Links

Some URL's of image search systems:
Altavista Picture Search: www.altavista.com/sites/search/simage
Lycos: http://lycos.com/picturethis/
PicToSeek: http://carol.wins.uva.nl/~gevers/PicToSeek/
QBIC: http://wwwqbic.almaden.ibm.com/
Hermitage-museum: http://www.hermitagemuseum.org/
BlobWorld: http://dlp.cs.berkeley.edu/photos/blobworld/
ImageRover: http://www.cs.bu.edu/groups/ivc/ImageRover/

5.5.4 Datamining for video retrieval

Alan Hanjalić [1]

With the objective of facilitating the interaction with large volumes of video material stored in emerging high-capacity digital video libraries, considerable scientific effort has been invested in the past decade in developing video-content analysis algorithms. As illustrated in Figure 1, these algorithms have the task of mining the raw digital video data in search of information related to the actual video content, which will be used in the subsequent video retrieval step where the retrieval requests, such as:

- Give me an overview of movie episodes.
- Find all news reports on Euro'.
- Let me see all highlights of the Olympic Games.
- I want all romantic movie scenes.
- Find me some funny action scenes.
- Play a short summary of a movie.

are handled. Just as in the case of image search, video-content analysis algorithms also operate on the signal characteristics of a digital video stream, that is, on low-level features. Since, however, the retrieval requests illustrated by the examples above are formulated at cognitive level and address the actual semantic content of a video, bridging the 'semantic gap' by finding models capable of establishing relations between low-level features and video semantics is the main challenge in developing video-content analysis algorithms.

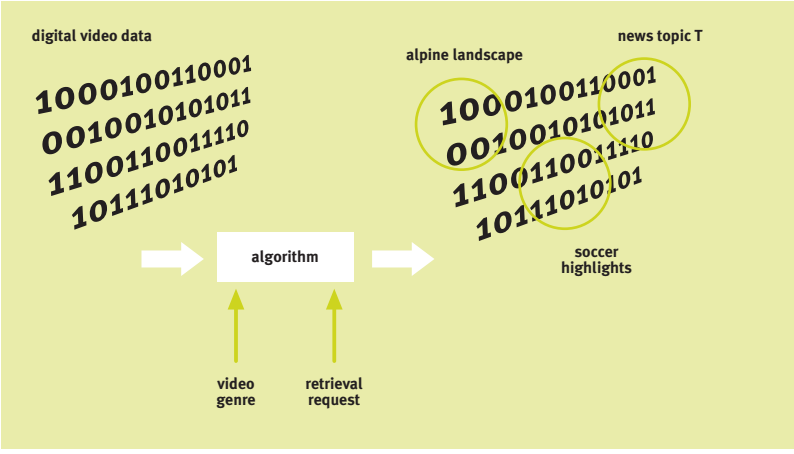


Figure 1
Mining raw digital video data in the search for information about video content.

Depending on which particular retrieval requests and video genres are considered, video-content analysis algorithms can be divided into five major categories:

1 Dr Ir A Hanjalić, A.Hanjalic@ITS.TUDELFT.NL, Delft University of Technology, Delft, The Netherlands

– video summarization;
– extraction of semantically meaningful segments from a video;
– semantics-based video classification;
– high-level segmentation;
– affective video-content extraction.

Video summarization

Since it aims at providing a first impression about a video to the user while keeping the information offered to the user as compact as possible, video summarization is an important step in providing efficient interaction with a large-scale video database. We refer here to video summarization as the process of concentrating the content of a long video into a limited number of frames or short video clips. As a result, the resulting frames and video clips show all relevant content elements of the original video, such as landscapes, objects, persons or situations. In this sense, a video summary is meant to be a rather objective short description of a video, suitable to provide a good insight into the video content to anyone. Recently proposed summarization approaches address this problem from various aspects [Gong, 2000a; Chiu, 2000; Toklu, 2000; Syeda-Mahmood, 2000; Gong, 2000b; Doulamis, 2000; Jeho, 1999; Uchihashi, 1999; Vaconcelos, 1998; Hanjalić, 1999a; Tiecheng, 2000; Sundaram, 2001].

Extraction of semantically meaningful segments

Extraction of semantically meaningful segments is basically a filtering of a video database and the isolation of the video segments that may be of interest for retrieval. Recently proposed algorithms belonging to this category are used to extract interesting events from a basketball game [Saur, 1997] and a soccer broadcast [Gong, 1995], to find commercial spots in various TV programs [Lienhart, 1997; McGee, 1999; Colombo, 1998], dialogs, actions and story units in a movie [Yeung, 1997] as well as movie trailers [Pfeiffer, 1996].
For instance, using the approach by Saur specific action scenes like wide-angle and close-up views, fast breaks, steals or potential scores may be extracted from a basketball game. In this way the user does not have to watch the entire game (which may also contain some boring parts that are not worth watching), but need only enjoy the extracted highlights. For instance, each video shot[2] is classified as wide-angle or a close-up shot by investigating the total motion intensity resulting from camera and object motion. While wide-angle shots are taken from a distance and are relatively stationary, close-up shots are highly dynamic, since the camera only shows a small portion of a scene and usually follows an object. In this way, close-up shots may be related to some exciting moments of the game. The term 'fast break' is defined as a 'fast' movement of the ball from one end of the court to the other. In order to detect fast breaks, one accumulates the magnitude of the motion vectors along a sequence in such

2 A video shot can be defined as a continuous camera drive, e.g. a zoom of a person talking or a sequence where a camera follows a car.

a way that the accumulation is reset to zero each time the motion changes direction. If the camera follows the ball during a fast break, a long and persistent pan is registered in these segments. Therefore, the search for fast breaks is actually the search for extremely long segments in the accumulation curve between two reset points. By exploring specific camera motion and lengths of corresponding video segments, one can also characterize steals and ball-possessions.

In the algorithm for analyzing soccer broadcast [Gong, 1995] the standard layout of a soccer field was used to classify the video material into nine different categories, examples of which are 'round the left penalty line' or 'near the top right corner'.

Commercials can efficiently be detected by investigating the shot-change frequency that can be assumed much higher in commercials compared to the rest of a program. A further possibility for extracting commercial breaks is the detection of an abrupt change in audio volume (pitch) that is considered to be a reliable indication about the commercials boundaries [Lienhart, 1997; McGee, 1999; Colombo, 1998].

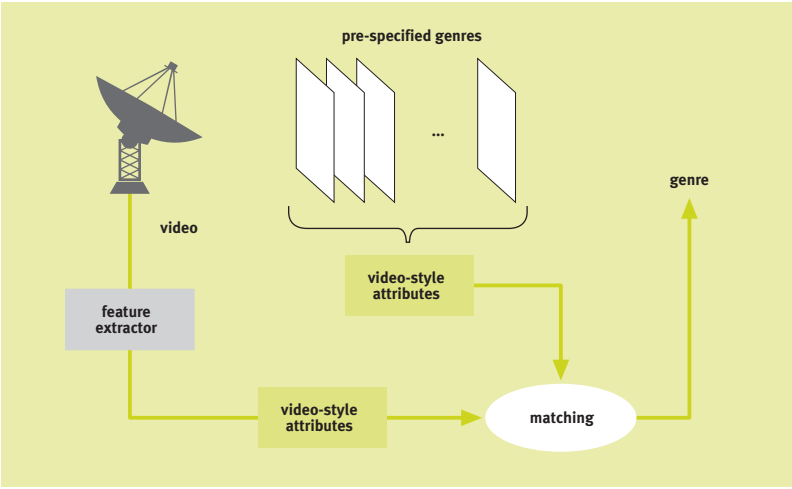
Yeung and Yeo propose in [Yeung, 1997] a method for recognizing semantic video segments by performing a 'semantic labeling' of video shots. This labeling is done by applying time constrained clustering to all shots in a sequence. Clustering is performed on the basis of visual similarity of shots and their mutual temporal locality. Visual similarity can be evaluated in various ways, such as comparing the color composition of characteristic frames (key frames) of the shots. Belonging to a certain cluster automatically assigns a label to a shot that corresponds to that cluster. Then, semantic segments are detected using the shot labels and their sequential order. For instance, a series of shots labeled as ABABABAB probably represents a dialog involving two characters. Also the high-action segments of a movie can be assumed at places where no or only a minimal repetition of shot labels is found, e.g. ABCDEFBGHIAK.

The extraction of most characteristic movie segments may be interesting for the purpose of automatically producing a movie trailer. As proposed in [Pfeiffer, 1996], movie segments to be included in such a trailer may be selected by investigating the specific visual and audio features and by taking those segments which are characterized by high motion (action), basic color composition similar to average color composition of a whole movie, dialog-like audio track, and high contrast. It is claimed that this method yields good quality movie trailers, since 'all important places of action are extracted', [Pfeiffer, 1996].

Semantics-based video classification

Semantics-based video classification is mostly performed in view of a number of pre-specified genres and aims at providing the top level of interaction between the user and a video database [Fischer, 1995; Huang, 2000;

Figure 2
Video genre classification scheme [Fischer, 1995].



Mallikarjuna, 1999; Huang, 1999; Girgensohn, 1999; Wang, 1997]. For instance, the method for detecting video genres recently proposed in [Fischer, 1995] consists of three steps (Figure 2). In the first step, the syntactic properties of a digital video, such as color statistics, shot-boundaries, motion vectors, simple object segmentation and audio-statistics, are analyzed. The results of the analysis are used in the second step to derive video-style attributes, such as shot lengths, camera panning and zooming, types of shot boundaries (abrupt ones vs. dissolves, fades, etc.), object motion and speech vs. music, which are considered to be the distinguishing properties for video genres. In the final step, an 'educated guess' is made about the genre to which the video belongs, based on a mapping of the extracted style attributes with those corresponding to different pre-specified genres. Experiments were reported using a number of sequences that were to be classified in one of the following genres: news, car races, tennis, commercials and animated cartoon. It is interesting to see in which way the style attributes were related to a particular genre. For instance, for a news program, the appearance of interchanging low- versus high-motion video segments is investigated. There, low-motion segments correspond to anchorperson shots, which are separated by high-motion report segments. Also, a distinction is made between the anchorperson and some other 'talking head' through the requirement that the periodically appearing low-motion segments need to be visually similar. This is done by computing and block-wise comparison of the histograms of three subsequent low-motion segments. On the other hand, tennis is a good example of how audio can be used for detecting a video genre. As reported by the authors, a tennis game has a highly pronounced structure of the audio stream, characterized by interchanging 'bouncing-ball' and speaker phases.

High-level segmentation

High-level segmentation is a content analysis step that is typical for video genres characterized by a clear sequential content structure. A video belonging to these genres can be modeled as concatenation of separate contexts – semantic segments – each of which is potentially interesting for retrieval. Examples of semantic segments are reports in a broadcast news program, episodes in movies, topic segments of documentary programs or scenes in a situation comedy. Semantic segments of a sequential video structure can be understood as concatenations of contextually related video shots. Consequently, the boundaries of shot segments are to be searched at shot boundaries obtained by performing the so-called low-level video analysis (Figure 3).

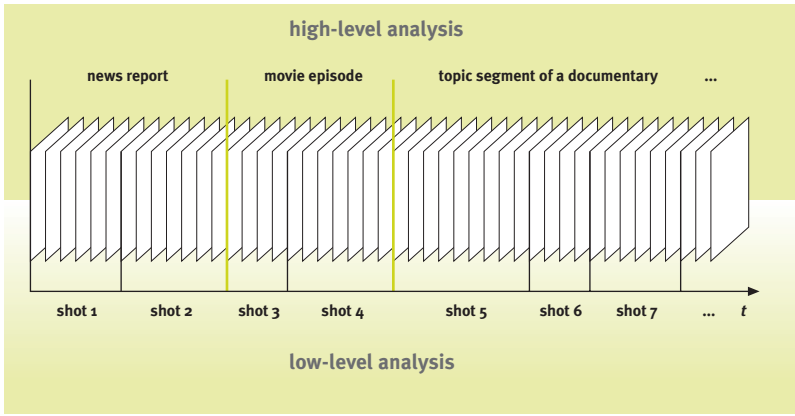


Figure 3
Low- versus high-level segmentation.

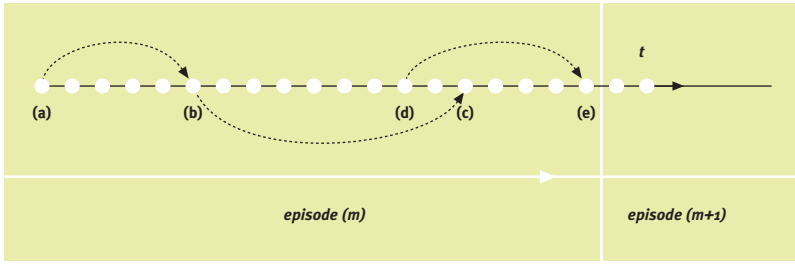


Figure 4
Illustration of the episode boundary detection procedure. The shots indicated by (a) and (b) can be linked and are by definition part of episode (m). Shot (c) is implicitly declared part of episode (m), since the shot (d) preceding (c) is linked to a future shot (e). Shot (e) is at the boundary of episode (m), since it cannot be linked to future shots, nor can any of its predecessors [Hanjalič, 1999b].

As described in [Hanjalić, 2001a], partitioning of a movie into episodes may be performed by measuring the coherence of the visual content over a longer series of shots. At places where this coherence is low, segment boundaries may be assumed. In [Hanjalić, 1999b] the coherence is modeled as a set of overlapping links that are established between shots containing certain (high) percentage of similar visual features. Then, a boundary between two neighboring semantic segments is assumed at the time stamp at which no further progression in establishing the overlapping links is possible. The underlying idea of the method is that a movie episode concentrates on an event or several overlapping events each of which is related to a certain scenery, people and objects. For this

reason, an episode is characterized by a global consistency of its visual content. As indicated by the links in Figure 4, this consistency is not necessarily present in two consecutive shots, but can be traced between two shots lying sufficiently close to each other.

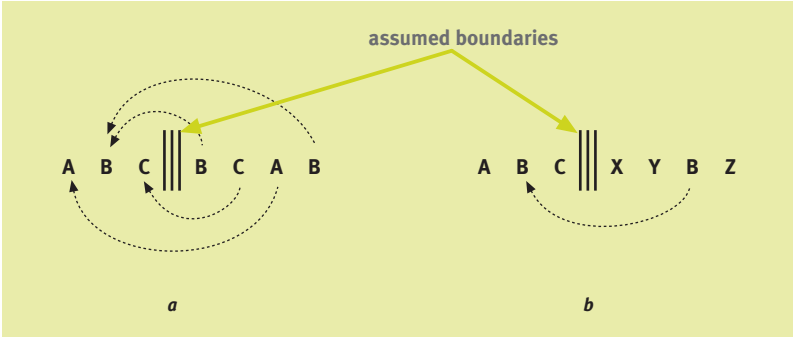


Figure 5
a *Good coherence (many future recalls),*
b *bad coherence (few future recalls) [Kender, 1998].*

Kender and Yeo [Kender, 1998] model the content coherence as a continuous function that is evaluated at each shot boundary. The model is based on the assumption that the more the present shot and its nearby successors remind the viewer of the prior shots, the higher is the coherence at the time stamp of the present shot. This is indicated in Figure 5a. Opposed to this, the coherence value is low, if the present shot and its nearby successors fail to remind the viewer of the previous content of a video sequence (Figure 5b). In view of the above, the coherence value is computed at each shot transition by checking the recalls of shots preceding the boundary by the shots following the boundary. How strong a recall between two shots is, depends on the similarity of their visual content and their lengths as well as on their relative temporal positions in a video sequence.

Affective video-content extraction

While the research efforts in providing reliable algorithmic solutions for the previous four categories have been invested continuously for almost a decade, first attempts to extract affective content of a video have been made only recently. The affective content of a video can be defined as the type and amount of feeling or emotion contained in and mediated by a video toward a user. The attribute 'affective' has emerged, on the one hand, from search requests that target the affect, that is, a feeling, mood or emotion. Examples of such requests are those for 'a romantic movie scene', 'glorious moments', 'funny action scenes' or 'happiness'. On the other hand, extracting this type of video content may also provide information about audience's affective response to mediated facts, or in other words, about how the audience feels while watching a video. Hanjalić and Xu proposed in [Hanjalić, 2001b] a method for quantifying the affective content of a video, that is underlined by the so-called 'dimensional

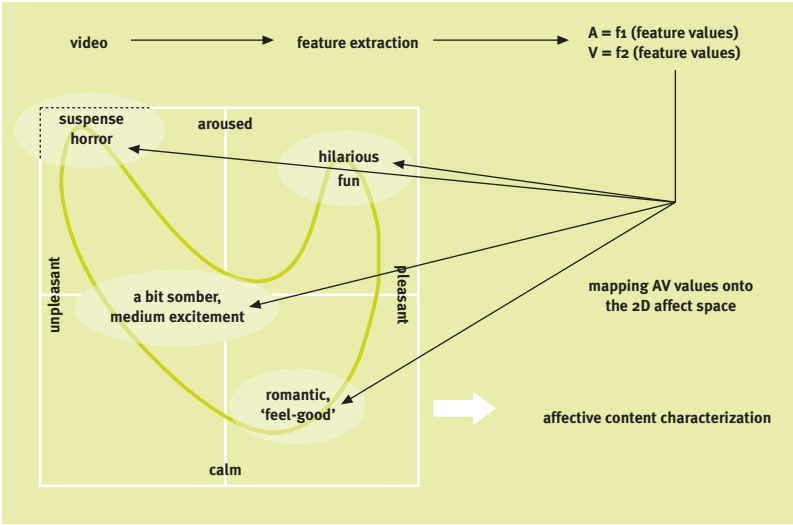
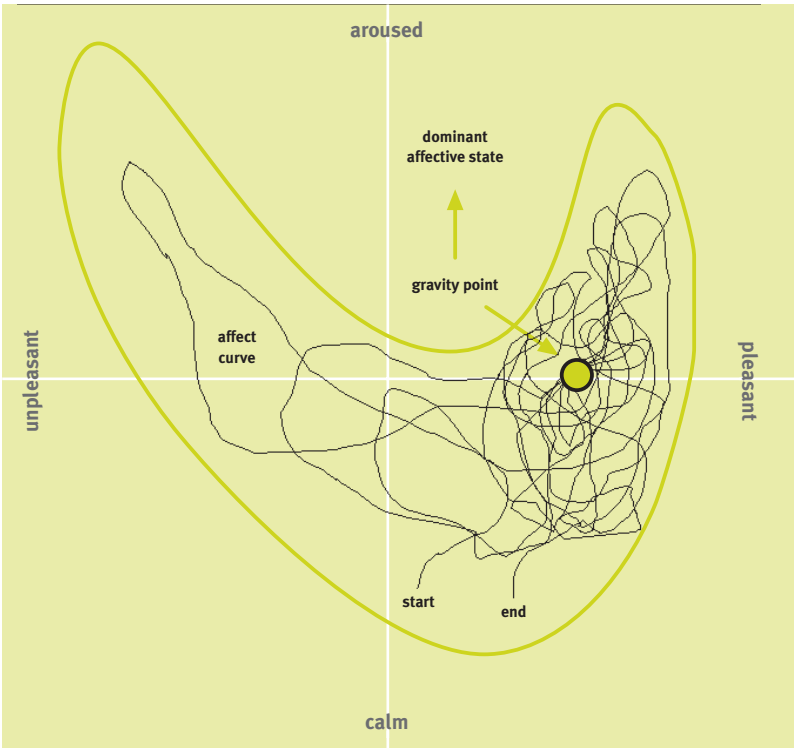


Figure 6
Measuring arousal and valence using suitable low-level features and mapping these values onto the 2D affect space [Hanjalić, 2001b].

approach to affect' known from psychophysiology studies. The method is based on obtaining the values of two affect dimensions (i) arousal – level of excitement – and (ii) valence – 'nature' of excitement, ranging from unpleasant to pleasant – by measuring video signal characteristics (low-level features) that are known to be related to each of these two affect dimensions. Every value pair of arousal and valence defines one affective state – a specific mood, feeling or emotion mediated by video toward the audience.

Figure 6 illustrates the mapping of arousal and valence values onto the so-called 2D affect space [Dietz, 1999], whose roughly parabolic shape circumvents the scattered plot of affective responses to a set of calibrated stimuli (IAPS [Lang, 1988], IADS [Bradley, 1991]) and which is therefore assumed to contain most of the common affective states. If one measures the arousal and valence values for consecutive time stamps along a video and marks the corresponding affective states in the 2D affect space, one can obtain the affect curve [Hanjalić, 2001b], as illustrated in Figure 7. The affect curve depicts the progress of a video through its various affective phases and also shows the transitions between audience's affective responses to mediated facts from one time stamp to another. Further, in order to extract parts of a video characterized by a certain mood or feeling, it is sufficient to select temporal video segments for which the affect curve passes through the corresponding areas of the 2D affect space. Finally, based on the position of the gravity point of the curve in the 2D affect space, the dominant affective state can be assigned to a video, providing in this way an additional video-content descriptor and a base for matching the content of a large video repository to the personal taste and interest of a user [Hanjalić, 2001b].

Figure 7
The affect curve [Hanjalič, 2001b].



Prospects

What once started as a modest effort at some industrial and academic research labs (IBM, MIT, ISS[3]) has grown into a research area involving an enormous number of people and showing its first concrete products (INFORMEDIA [Christel, 1994] and VIRAGE [Bach, 1996]). We can anticipate that further development of video-content analysis algorithms will strongly accelerate in the years to come. This is not only because of an enormous scientific challenge stimulating an army of scientists all over the world, but rather in view of a serious problem that emerges in all spheres of our lives — multimedia data overload — where video plays the most important role, both regarding huge amounts of data related to it and its popularity compared to other media. For this reason the development of video-content analysis mechanisms can indeed be considered as one of the most important data mining-related trends in the years to come. Nobody dares to claim that the semantic gap will be bridged in 20 years. However, we can expect the theory and tools that facilitate video retrieval and management to move a further big step toward maturity.

3 Institute for Systems Science, National University Singapore.

References

- Bach, J.R., et al. (1996). The Virage Image Search Engine: An Open Framework for Image Management. Proceedings of IS&T/SPIE Storage and Retrieval for Still Image and Video Databases IV. Volume **2670**
- Bradley, M.M., P.J. Lang. (1991). International Affective Digitized Sounds (IADS): Technical Manual and Affective Ratings. Gainesville. University of Florida, Center for Research in Psychophysiology
- Chiu, P., A. Girgensohn, W. Polak, E. Rieffel, L. Wilcox. (2000). A Genetic Algorithm for Video Segmentation and Summarization. Proceedings IEEE International Conference on Multimedia and Expo 2000 (ICME 2000). Volume **3**. pp329-1332
- Christel, M., S. Stevens, H. Wactlar. (1994). Informedia Digital Video Library. Proceedings of ACM 2nd International Conference on Multimedia. Video Program. New York
- Colombo, C., A. del Bimbo, P. Pala. (1998). Retrieval of Commercials by Video Semantics. Proceedings IEEE Conference on Computer Vision and Pattern Recognition 1998. Volume **2**. pp572-577
- Dietz, R., A. Lang. (1999). Affective Agents: Effects of Agent Affect on Arousal, Attention, Liking and Learning. http://www.polara.org/pubs/cogtech1999.html
- Doulamis, A.D., N.D. Doulamis, S.D. Kollias. (2000). Efficient Video Summarization Based on a Fuzzy Video Content Representation. Proceedings IEEE International Symposium on Circuits and Systems, 2000 (ISCAS 2000). Volume **4**. pp301–304
- Fischer, S., R. Lienhart, W. Effelsberg. (1995). Automatic Recognition of Film Genres. Proceedings ACM Multimedia '95. San Francisco
- Girgensohn, A., J. Foote. (1999). Video Classification Using Transform Coefficients. Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing 1999. Volume **6**. pp3045-3048
- Gong, Y., et al. (1995). Automatic Parsing of TV Soccer Programs. Proceedings IEEE International Conference on Multimedia Computing and Systems
- Gong, Y., X. Liu. (2000a). Generating Optimal Video Summaries. Proceedings IEEE International Conference on Multimedia and Expo 2000 (ICME 2000). Volume **3**. pp1559-1562
- Gong, Y., X. Liu. (2000b) Video Summarization Using Singular Value Decomposition. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition 2000. Volume **2**. pp174-180
- Hanjalić, A., H. Zhang. (1999a). An Integrated Scheme for Automated Video Abstraction Based on Unsupervised Cluster-Validity Analysis. IEEE Transactions on Circuits and Systems for Video Technology. Special Issue on Object-Based Video Coding and Description

- Hanjalić, A., R.L. Lagendijk, J. Biemond. (1999b). Automated High-Level Movie Segmentation for Advanced Video Retrieval Systems. IEEE Transactions on Circuits and Systems for Video Technology
- Hanjalić, A., R.L. Lagendijk, J. Biemond. (2001a). Recent Advances in video-content analysis: From Visual Features to Semantic Video Segments. International Journal of Image and Graphics **1** (1). World Scientific, Singapore
- Hanjalić, A., L.-Q. Xu. (2001b). User-Oriented Affective Video Analysis. IEEE Workshop on Content-based Access of Image and Video Libraries, in conjunction with the IEEE CVPR 2001 Conference. Kauai, Hawaii (USA)
- Huang, J., Z. Liu, Y. Wang, Y. Chen, E.K. Wong. (1999). Integration of Multimodal Features for Video Scene Classification Based on HMM. IEEE 3rd Workshop on Multimedia Signal Processing. pp53–58
- Huang, J., Z. Liu, Y. Wang. (2000). Joint Video Scene Segmentation and Classification Based on Hidden Markov Model. Proceeding IEEE International Conference on Multimedia and Expo (ICME 2000). Volume **3**. pp1551-1554
- Jeho, N., A.H. Tewfik, Video Abstract of Video. (1999). Proceedings IEEE 3rd Workshop on Multimedia Signal Processing. pp117–122
- Kender, J.R., B.-L. Yeo. (1998). Video Scene Segmentation via Continuous Video Coherence. Proceedings IEEE Conference on Computer Vision and Pattern Recognition. Santa Barbara
- Lang, P., J.M.K. Greenwald. (1985). The International Affective Picture System Slides and Technical Report. Gainesville, University of Florida, Center for Research in Psychophysiology
- Lienhart, R., C. Kuhmuench, W. Effelsberg. (1997). On the Detection and Recognition of Television Commercials. Proceedings IEEE ICMCS '97. Ottawa, Canada
- Mallikarjuna, R.K., K.R. Ramakrishnan, N. Balakrishnan, S.H. Srinivasan. (1999). Neural Net Based Scene Change Detection for Video Classification. IEEE 3rd Workshop on Multimedia Signal Processing. pp247–252
- McGee, T., N. Dimitrova. (1999). Parsing TV Programs for Identification and Removal of Nonstory Segments. Proceedings IS&T/SPIE Storage and Retrieval for Image and Video Databases VII. Volume **3656**
- Pfeiffer, S., R. Lienhart, S. Fischer, W. Effelsberg. (1996). Abstracting Digital Movies Automatically. Journal of Visual Communication and Image Representation **7** (4):345-353
- Saur, D.D., Y.-P. Tan, S.R. Kulkarni, P.J. Ramadge. (1997). Automated Analysis and Annotation of Basketball Video. Proceedings of IS&T/SPIE. Volume **3022**
- Sundaram, H., S.-F. Chang. (2001). Constrained Utility Maximization for Generating Visual Skims. IEEE Workshop on Content-Based Access of Image and Video Libraries, in conjunction with the IEEE CVPR 2001 Conference. Kauai, Hawaii (USA)

- Syeda-Mahmood, T., S. Srinivasan, A. Amir, D. Ponceleon, B. Blanchard, D. Petkovic. (2000). CueVideo: a System for Cross-Modal Search and Browse of Video Databases. Proceedings IEEE Conference on Computer Vision and Pattern Recognition 2000. Volume **2**. pp786-787
- Tiecheng, L., J.R. Kender. (2000). A Hidden Markov Model Approach to the Structure of Documentaries. Proceedings IEEE Workshop on Content-Based Access of Image and Video Libraries. pp111–115
- Toklu, C., S.-P. Liou, M. Das. (2000). Video Abstract: a Hybrid Approach to Generate Semantically Meaningful Video Summaries. Proceedings IEEE International Conference on Multimedia and Expo 2000 (ICME 2000). Volume **3**. pp1333-1336
- Uchihashi, S., J. Foote, Summarizing Video Using a Shot Importance Measure and a Frame-Packing Algorithm. (1999). Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing 1999. Volume **6**. pp3041-3044
- Vasconcelos, N., A. Lippman. (1998). A Spatiotemporal Motion Model for Video Summarization. Proceedings IEEE Conference on Computer Vision and Pattern Recognition. pp361–366
- Wang, Y., J. Huang, Z. Liu, T. Chen. (1997). Multimedia Content Classification Using Motion and Audio Information. Proceedings IEEE International Symposium on Circuits and Systems (ISCAS '97). Volume **2**. pp1488-1491
- Yeung, M., B.-L. Yeo. (1997). Video Visualisation for Compact Presentation and Fast Browsing of Pictorial Content, IEEE Transactions of Circuits and Systems for Video Technology. Special Issue on Multimedia Technology, Systems and Applications

5

5.6 Web Mining

Raymond Kosala[1], *Hendrik Blockeel*[2], *Frank Neven*[3]

5.6.1 An overview of web mining

Introduction

The World Wide Web is an enormous information repository, but in order to efficiently access the information contained in it, sophisticated software is needed. This software is to a large extent based on data mining technology. In this section we start with a brief discussion of the nature of the Web and the specific problems encountered when trying to find information. We then give an overview of the many techniques that exist for web mining, illustrated with applications of these techniques. Note that with this overview we aim only at a high-level discussion of techniques; for a more thorough discussion we refer to [Kosala, 2000].

1 Ir R. Kosala, Raymondus.Kosala@cs.kuleuven.ac.be, Katholieke Universiteit Leuven, Department of Computer Science, Leuven, Belgium, http://www.cs.kuleuven.ac.be/~raymond/

2 Dr H. Blockeel, Hendrik.Blockeel@cs.kuleuven.ac.be, Katholieke Universiteit Leuven, Department of Computer Science, Leuven, Belgium, http://www.cs.kuleuven.ac.be/~hendrik/

3 Prof Dr F. Neven, frank.neven@luc.ac.be, University of Limburg (LUC), Department WNI, Infolab, Leuven, Belgium, http://alpha.luc.ac.be/~fneven/

The World Wide Web

In the early nineties, people started to use the term 'World Wide Web' (WWW) to refer to the rapidly growing network of computers that were connected to each other via the Internet. While the Internet itself is much older, it is only around that time that the concept of the Web as a single, huge and distributed repository of information came into existence. Several evolutions contributed to this:

- More standardization started to appear in the way in which information was made available (the concepts of web pages and web sites evolved).
- The use of HTML created a separation between the logical format and the physical format of the stored information, making the physical format transparent to the user. Instead of needing to search through the directories of some file system, looking for files of a certain type, the structure of the information repository was reflected in so-called hypermedia: easily readable documents that contain text, images, etc., but most of all contain links that connect to other documents.
- The concepts of Internet nodes and connections became largely invisible for the user. By clicking on a word in a hypertext, the user 'follows a link to another document'. It is irrelevant whether the other document resides on the same computer or another one; from the user's point of view it is as if all the information is simply available on their own computer.
- Sites appeared that indexed the mass of information on the Web, and thus could serve as entry points ('portals') for users looking for specific information. These sites typically also offer search engines.

Thanks to these developments the Web can now be seen as a gigantic database that contains many different kinds of information, offering many different ways to query. Moreover, anyone can contribute to this database, not only by providing new information on their own web site, but also by extending the querying possibilities of the Web. For instance, anyone who feels they have found a better way of accessing the information on the Web could in principle just write an interface and 'plug it in' by putting the interface on their own web site, thus extending the technical capabilities of the Web.
In this sense the WWW can be considered, if not the largest, certainly the most flexible database ever to have existed.

Finding information on the Web

Unfortunately, there is a drawback to the Web as an information repository. While the quantity of information that is — potentially — available is huge, in practice it may be much less, in the sense that it may not be obvious how to obtain it. If we define accessibility of information as the ease by which the information can be obtained (i.e. how much knowledge or expertise is needed to obtain it), then the accessibility of information on the Web leaves much to be

desired. The knowledge and expertise involved in getting certain information from the Web (in addition to being able to use a web browser) may consist of:

– knowing on which URL [4] the information resides;
– knowing where to find and how to use a search engine on the Web; the use of a search engine may range from very simple queries (entering a keyword) to rather complex queries (involving Boolean operators, or the specification of a domain name);
– knowing how to write a special-purpose agent that searches the Web for the information.

It is easy to find information, if one knows where it is, but it is unreasonable to assume that the user always has this information. On the other hand, the third option (writing a special-purpose program) presupposes a level of expertise that very few users have. Hence, the availability of good and easy-to-use search engines is crucial for increasing the accessibility of information on the Web. The reason why we consider the information on the Web to be poorly accessible is that the quality of current-day search engines still leaves much room for improvement.

Certain kinds of information can easily be found by using a search engine. For instance, if the user types in 'Amsterdam' as a keyword to search for, the chance is very high that the first pages the search engine returns contain information about the city of Amsterdam. But if the user wants to learn about computers, typing in 'computers' may give you millions of pages, only very few of which will contain relevant information. Typing in a phrase such as 'learn about computers' will probably help, but may still return too many pages, the majority of which are of little interest; moreover the most interesting ones may be so far down the list that the user will never discover them, and some very interesting pages may not even be in the list, because they do not contain the keywords themselves but only related words. Search quality can be quantified with two parameters: recall and precision, explained in Inset 1.

Inset 1: Recall and precision

The terms 'recall' and 'precision' are often used in the domains of information retrieval and extraction to indicate the quality of the result of a search. In the context of the above example, 'recall' refers to the proportion of the actually interesting pages that the search engine returns, and 'precision' refers to the proportion of pages returned by the search engine that are actually interesting.

More formally: given a set S (e.g. the set of all web pages) from which we want to extract the set of all members of S that satisfy some criterion C (denoted S[C]), and the result of the extraction process is a set A, then:

Recall R = |A[C]| / |S[C]|

Precision as P = |A[C]| / |A|

4 Universal Resource Locator, internet address.

In the above example S is the set of all web pages on the Web; however in other contexts it could be the set of all web sites (where a site is defined as a collection of pages that all belong together), the set of all images occurring on the Web, the set of all words occurring on the Web, the set of all words occurring in a given document, etc.

Obviously, when looking for information on the Web, the user prefers answer sets with high recall and precision; ideally the system should return everything that is of interest, and nothing else. Present day web technology is limited in the sense that for many kinds of questions, it is very hard to formulate a question in such a way that a set of answers with high recall and precision is returned[5].
In order to improve the situation, there is a need for more intelligent search engines. Several research domains are involved in building systems that return relevant information from a large information repository. We can distinguish:

- Information retrieval (IR): in this research domain the question of how to find relevant documents is studied. This problem is closely related to the problem of finding web pages that was discussed above.
- Information extraction (IE): in this research domain study centers on how to extract certain information from a single document. Assuming the document is an article published on the Web, finding the name of the author of the article or finding the names of all authors mentioned in the references section of the article are typical information extraction tasks.

These research domains are linked to each other by the technology they use, and in the same way they are linked to knowledge discovery and data mining. Indeed, techniques from data mining can be used for the goals of information retrieval and information extraction, as we shall see later on. However, because of its ability to construct new knowledge, data mining (or the knowledge discovery process as a whole) can also be used for yet another task: extending the Web with new information. This information could itself be made public on the Web, or it could serve the private purposes of the user (for instance, supporting knowledge-based inference and problem solving, see Example 2). When knowledge discovery is a goal in itself, then information retrieval and extraction become a preceding process, because standard data mining techniques usually cannot work with such heterogeneous information as is found on the Web, and preprocessing the data may involve IE and IR technology.
In summary, there is a complex interaction between data mining and information retrieval/extraction: both may employ the other to achieve their goals. A data mining technique might use data that have been extracted from the Web using IR and IE, and the latter may again use (other) data mining techniques.
In accordance with [Etzioni, 1996] we define web mining as the whole of data mining and related techniques that are used to automatically discover and

5 Actually recent research [Lawrence, 1999] has shown that a large percentage of the Web is not even indexed, which further reduces the maximum possible recall of search engines. The problem of indexing is out of the scope of this text; here we focus on obtaining information from those parts of the Web that are indexed.

extract information from web documents and services. We add 'and related techniques' here, because the term 'mining' in this context is often used in a more general sense than as only referring to data mining in the classical sense.

Data mining technology can help in many different ways to improve the intelligence of search engines. In the following we distinguish three different approaches:

1 Web content mining: investigating the content of documents in order to find relevant information.
2 Web structure mining: using the structure of the Web (i.e. the way in which different web pages are linked together) to find relevant information.
3 Web usage mining: using previously stored knowledge about the behavior of human users of the Web (for instance, how they navigated through the Web) to find relevant information.

In all three cases, the focus is on the use of data mining techniques for retrieval and extraction of information that is already there somewhere on the Web, not on the construction of new knowledge (which is a separate goal). In the following three sections we will give an overview of these different approaches. Next we discuss knowledge discovery and information integration, and finally we conclude with some practical illustrations.

Web content mining

Web content mining concerns the use of data mining techniques on the level of the contents of web documents. We distinguish two views here: the database view, where the Web is more or less considered to be a normal, relatively structured database which can be queried using some structured query language; and the information retrieval (IR) view, where the Web is considered a collection of largely unstructured documents.
In this part of the text we focus mainly on the mining of structured or text documents. Techniques for mining information hidden inside images, audio files, etc., which we refer to as multimedia mining, are the subject of Chapter 5.5.

The database view

HTML

In the database view, web documents are considered to be much more structured than in the IR view. Documents on the Web are defined in HyperText Markup Language (HTML). When mining inside an HTML document, the structure of the document as indicated by the HTML tags will be exploited.
However, the structure imposed by HTML is purely for presentation purposes. Indeed, HTML only provides tags to specify the title of the document, to parti-

tion the document into paragraphs, to indicate lists, tables, hyperlinks, and so on. The HTML file in Inset 2, for instance, displays the page in Figure 1 and could be part of a web page of a computer vendor where each separate page contains the data of each offered computer.

Inset 2: HTML

Tags are the words between brackets and determine how the text in between should be displayed. For instance, the text Laptop 44X3D between the start-tag <TITLE> and the end-tag </TITLE> is the text displayed in the title bar of the web browser. Two matching tags together with the text in between is called an element. Further, <BODY> specifies the content of the HTML file and each <LI> determines a list item. An example of a HTML file is given below.

```
<HTML>
   <HEAD>
      <TITLE> Laptop 44X3D </TITLE>
   </HEAD>
   <BODY>
      <LI> ABT </LI>
      <LI> grey </LI>
      <LI> 2000 </LI>
   </BODY>
</HTML>
```

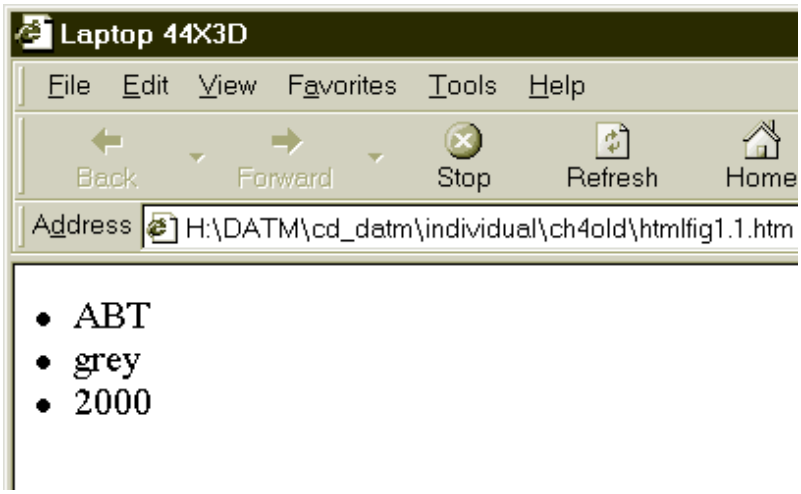


Figure 1
Screen grab of browser displaying Inset 2.

Although HTML, based on tags, is an excellent mechanism to provide platform independent browsing, it hardly imposes any semantics. Clearly, ABT, grey, and 2000 are properties of the laptop 44X3D and a human can defer their meaning, but not a computer program. Additionally, it could be possible that different lap-

top models have different properties and there is no way to specify this in HTML, while keeping the structure and the content of the document separated.

XML

Currently there is ongoing work in the area of what is called a 'semantic web' [Berners-Lee, 1998]. The idea here is to make the Web more understandable to computers by providing semantic tags in documents. An important impulse in this direction is given by the use of XML (eXtensible Markup Language) instead of HTML for web documents. XML is a new standard[6] for the specification of structured documents developed by the World Wide Web Consortium (W3C) and is essentially a cleaned up version of the Standard Generalized Markup Language (SGML). However, for the purpose of this section we can say that XML is just HTML with user definable tags.

Like HTML, XML adds extra information by means of tags. Only in the latter case, this information is not longer restricted to presentation (see Inset 3). For instance, the tag <supplier> indicates that ABT is the supplier of the laptop model 44X3D. Hence, every application that is capable of reading XML 'knows' that this vendor offers a grey laptop 44X3D of ABT at the price of 2000.

Inset 3: XML

The information in Figure 1, for instance, could be represented in XML as follows:

```
<product>
   <model> Laptop 44X3D </model>
   <supplier> ABT </supplier>
   <color> grey </color>
   <price> 2000 </price>
</product>
```

In comparison with relational databases, tags perform the function of column names of relations. XML, however, has the advantage that it can deal with irregularly structured data. Furthermore, it is platform independent and even application independent in the sense that one does not need the program that has generated the XML file as it is just ASCII. Maybe the most important advantage is that almost any data format can be readily translated into XML, which makes it suited as an intermediate format for data exchange on the Internet. In fact, many software vendors already bet on XML to become tomorrow's universal data exchange format and build tools for importing and exporting XML documents.

6 Included on the CD-rom XML in 10 points, XML 1.0Second Edition, URL: http://www.w3.org/XML

DTD's

As the tags in the XML document describe its semantics, XML is often called 'self-describing'. Nevertheless, for information extraction and integration purposes, it is convenient to have some information in advance on the structure of a collection XML documents. Such information is provided by Document Type Definitions (DTD's), which are essentially grammars. In brief, a DTD specifies the regular expression pattern for every element that subelement sequences of the element need to conform to. A document that conforms to a specific DTD is said to be valid with respect to this DTD. For the above example, such a DTD could say that each XML file consists of a sequence of products, where each product consists of a model, a supplier, a color, and a price, that the order does not matter and that, for instance, no tag is compulsory. In general, the structure of a document can be quite complicated as elements can be arbitrarily nested.
A document's DTD, hence, serves the role of a schema specifying the internal structure of the document. DTD's are critical to realizing the promise of XML as the data representation format that enables free interchange and integration of electronic data. Indeed, without a DTD, tagged documents have little meaning. Moreover, once major software vendors and corporations agree on domain-specific standards for DTD formats, it would be possible for inter-operating applications to extract, interpret, and analyze the content of a document based on the DTD it conforms to. Despite their importance DTD's are not mandatory and an XML document may not have an accompanying DTD. This may be the case, for instance, when large volumes of XML documents are automatically generated from, say, relational data, flat files, or semi-structured repositories. Therefore, it is important to build tools that infer schema information from large collections of XML documents. Garofalakis for instance, developed the tool XTRACT for inferring DTD's [Garofalakis, 2000]. However, to overcome the limited typing capabilities of DTD's, a lot of other formalisms, like XML schema, XDR, SOX, Schematron, DSD, and RELAX, are currently being developed, but none of them is a standard yet. Hence, a lot of work remains to be done in this area.

Queries on XML data

When schema information is present, information extraction reduces to writing queries in an XML query language. Although there is at the moment no standard XML query language, several have emerged over the past two years. Some of them, like XML-QL [Deutsch, 1999] and Lorel [Abiteboul, 1997] are based on query languages developed for semi-structured data. In brief, they consist of a WHERE and a CONSTRUCT clause. The WHERE clause selects parts of the input document, mainly by means of a pattern with variables, while the CONSTRUCT clause determines how these selected parts should be assembled to form the output. Consider, for instance, the XML-QL query given in Inset 4.

Inset 4: XML-QL query

```
WHERE <product>
            <model> $m </model>
            <supplier> ABT </supplier>
          </product>
CONSTRUCT <ABT>
                  <supplies> $m </supplies>
              </ABT>
```

The WHERE clause selects all models occurring in a <product> element supplied by ABT. Here $m is a variable. The CONSTRUCT clause specifies that for each match for $m an element <ABT> should be created with a subelement <supplies>. XML-QL has also more advanced features like tag variables, subqueries, aggregates, and path expressions.

Other languages

XSL[7] is a template based language developed by W3C. Initially, the aim of this language was to support easy transformations from XML to HTML. However, recent additions lifted XSL to a full fledged XML transformation language. Although XSL is definitely not a query language in the usual sense, as it is much too procedural and too difficult to use, it is the only one commercially available. Another language is XML-GL[8], which is graphical and therefore well-suited for supporting a user-friendly interface. Finally, we mention the language Quilt [Robie, 2000] which is a mixture of features of XSL and declarative WHERE and CONSTRUCT constructs.

As XML is a relatively new topic, not all these languages are already well studied (but see [Bonifati, 2000; Bex, 2000]) and it remains to be seen which language and which features will make it into a standard.
For more information on XML, XML query languages, and semi-structured data, we refer to [Abiteboul, 1999].

The information retrieval view

In the IR view, web documents are viewed as poorly structured resources, hence these approaches do not depend much (or at all) on such structure. For instance, XML tags are typically not available, and HTML tags may be available but sparse; hence the system should not depend on them too much. Instead, IR approaches depend more on statistical properties of documents (e.g. frequencies of specific words), or on grammatical analysis (deep or shallow) of text. The properties of documents that can be used depend on the representation of the documents. An often used representation is the 'bag of words' representation, where a document is described by listing how many times each word

7 http://www.w3.org/Style/XSL

8 http://xerox.elet.polimi.it/Xml-gl/index.html

occurs in it (the position of the words is thus ignored). A document would for instance be considered highly relevant for a keyword search if the keyword occurs often (relative to its 'average' occurrence in documents in general).
This representation can be improved by finding and exploiting correlations between words. A simple technique is stemming (e.g. 'cat' and 'cats' could both be counted as occurrences of the stem 'cat'), more advanced techniques may try to find related topics and synonyms by analyzing the co-occurrence of words in documents, etc. Instead of words, one may also use n-grams (word sequences of length n) or phrases; or one can add (limited or full) information about the positions of the words. Finally, an interesting line of research is that of topic detection and tracking (TDT) [Allan, 1998], where the aim is to follow a 'story', a thread of related events, throughout several documents. Thus the context of a single document (its place in the story) can be used to obtain information on what it is about.

Several interesting applications of the 'information retrieval' type of web mining exist. We mention Personal WebWatcher [Mladenic, 1996][9], which is a web browsing assistant that accompanies the user, when browsing from page to page and highlights interesting hyperlinks. This system generates a user profile based on the content analysis of the requested web pages without soliciting any keywords or explicit rating from the user. Another example is NewsWeeder [Lang, 1995], an intelligent agent that filters electronic news. The system has a common user interface that enables the user to search and access the news, and provides some additional functionalities to collect user's ratings as feedback. From this information, the NewsWeeder assigns the predicted relevance of each article with text classification methods and generates a list of the top articles found according to the user profile.
Another example of a method of the information retrieval type is given in Section 5.6.2, Extracting knowledge from the Web.

Web structure mining

Web structure mining denotes the use of data mining techniques on data about the structure of web sites, i.e. the way in which different pages are linked to each other. This structure can be investigated on several levels: locally on a web site (i.e. the way in which pages on the same web site are linked together), or more globally by also taking into account pages on other sites that are linked (possibly indirectly, through a chain of links) to the page under investigation. The structure of the Web is usually modeled as a graph, which is a natural way to represent the connectivity of the links in the WWW.
Web structure mining can be useful for a variety of tasks. Common tasks are finding interesting sites, web communities and topics.

9 See also Machine learning on distributed text data, Mladenic's PHD thesis on the CD-rom.

Interesting sites, authorities and hubs

The first task we will discuss is the identification of web sites that are of high general interest (i.e. not necessarily just for some specific topic). This approach is based on the assumption that a link to another site can very often be viewed as an implicit endorsement. For example, many personal web sites direct people to Yahoo as a search engine; this can be viewed as an indication that Yahoo is an interesting site. Of course many links are just for navigational purposes ('return to the main page'), they can be advertisements, or even point to a site explicitly disapproved of; but when a large enough number of links is present the number of such 'false' links is usually negligible and hence the existence of many links to a web page can be seen as an indication of authority [Kleinberg, 1998]. When a web site provides many links to other popular web sites, we can call it a hub. Thus, web structure mining is useful to find authority sites and hubs.

Web communities

Next to identifying interesting sites, web structure mining can also be used to discover entities at a higher level, e.g. collections of sites [Kumar, 1998]. One example of this is a so-called 'web community': a collection of web sites that have similar contents and aim at users with similar interests, and that are usually highly interlinked. One could consider, for instance, a collection of soccer fan sites in which most sites have links to most other fan sites. A user wanting to find information on soccer will be helped better, if the existence of this collection of sites is explicitly mentioned in the result of a query than if all these sites occur as different answers. It is obvious that web structure mining can play a major role in the discovery of web communities.

Topics

A third application of web structure mining is in learning which topics are related to each other. Using web content mining, one could infer that two different words are somehow related to each other, if they often co-occur in the same document. With web structure mining the same can be done on a more global level: one can, e.g. infer that two different topics are related, if many links exist between pages that are about topic 1 and pages about topic 2, even though the two topics do not often co-occur on the same page.

Combined structure and content mining

The tasks discussed above are examples of global web structure mining. Similar techniques can be used on a more local level, although in this case they are often highly intertwined with web content mining and the difference between both is not always clear-cut (as pointed out in [Kosala, 2000], where the term 'web structure mining' refers mainly to structure mining on a more global level). A nice application of combining content and structure mining is motivated by

the idea that it is often easier to derive what a page is about by looking at the contents of both the page itself and the contents of pages that have links to the page. [Blum, 1998] follow this approach in the context of their study of co-learning. One learning task they consider is learning to tell whether a web page is the home page of a course or not. In order to classify pages, they not only look at words occurring on the page itself, but also at the words associated with the links on other pages that point to the page. (E.g. the page itself may not contain the phrase 'home page' but a student might link to the page with the words 'home page of data mining course'; the latter is indeed a strong hint that the page is indeed a course home page). Blum and Mitchell showed that their approach of looking also at other pages increased the classification accuracy of their system significantly.

Web usage mining

The idea behind web usage mining is to infer information from the behavior of users of the Web. Web usage mining is often quite visible on the Web. A typical example is cross-selling, a functionality offered by many on-line stores: when the user asks information about a book, the site offers, among other things, information of the kind "people who bought this book also bought..." In this way the store can point the user to other potentially interesting information, bypassing any investigation of contents or structure of web documents, but instead using a log of the behavior of users of the site. It will be clear that in this way connections can be found that would be hard to find using content or structure mining. Instead, the intelligence that humans demonstrate in selecting the right information is directly exploited by the computer system. This is an important and fundamental advantage of web usage mining over the other web mining categories we have discussed.

The applicability of web usage mining is much broader than the above example suggests, and it can also be applied in less visible ways. For instance, a search engine that keeps logs of the keywords typed in by its users might detect that certain keywords often occur together or are used for successive searches. In this way a search engine might try to assist the user by reporting on pages that do not contain the exact words typed in by the user, but contain related words, or it might at least suggest to the user to try searches containing those words.

Self-adapting web pages are another interesting example of web usage mining. Such web pages keep logs of how many times users follow certain links, and based on this information rearrange themselves so that the links that the page thinks are of most interest to the current user are made more clearly visible. More on adaptive web sites can be found in Section 5.6.3, Mining for adaptive web sites.

Web usage mining is getting more important as more and more companies adopt e-business strategies. Corporate web sites are becoming new channels

for customer relationship management (CRM). A tremendous amount of web usage data is generated by the interactions of customers with the Web, such as cookies, form data, referrer data, server log data, session data, etc. These data (possibly combined with content and structure data) could become a source of valuable knowledge about web users or customers. Information providers could learn about customer interests, preferences, life style, and behavior individually or collectively. With web usage mining, individual CRM and marketing could be done intelligently. Learning about customers and matching their preferences provides a new competitive advantage.

Knowledge discovery and information integration

While the previous sections discussed the use of data mining techniques for extraction of already existing knowledge, the subject of this section is mainly the construction of new knowledge. Classically, knowledge discovery focuses on finding patterns or relationships in data sets. However, in the area of web mining another approach deserves attention, one that we refer to as 'information integration'. Here the focus is not so much on finding patterns, but on combining existing chunks of information into one consistent whole. The difficulty of this task arises from the fact that information on a single topic is kept at several places on the Web, and possibly also in different formats. This issue is not typically raised in classical knowledge discovery approaches, although it may gain importance there too.

Indeed, as web data are becoming an increasingly important component in making business decisions, more and more companies recognize the need to use web data to extract additional value. Besides the external data such as those available from the Web, even today's companies' internal data sources are heterogeneous and distributed in different locations to accommodate different needs and types of customers. This situation creates problems, when one wants to derive a single, comprehensive view of these data. Information integration can be seen as an automated method for querying across multiple heterogeneous data sources in a uniform way. Once the data have been integrated, they are more easily accessible for further analysis by either human analysts or data mining systems.
Basically there are two approaches to web information integration. These are the virtual database [10] and the web warehousing approach [11].

Warehousing approach

In the warehousing approach, data from some web sources is collected into a data warehouse and all queries are applied to the warehouse. The advantage of this approach is that performance can be guaranteed at query time. The disadvantage is that the data in the warehouse could be outdated, if not checked frequently.

10 For example http://www.jango.com

11 For example http://www.junglee.com

Virtual database approach

In the virtual approach, the data are not replicated, but stay in their original locations. All queries are posed to a mediated or global schema, usually designed for a specific application, inside a mediator. The mediated or global schema is a set of virtual relations that are not actually stored anywhere. The mediator then translates the queries from users into queries that refer directly to the source schemas, usually done through wrappers. The advantage of the approach is that the data is guaranteed to be fresh. The disadvantage is that even with sophisticated query optimization and execution mechanisms, good performance is not guaranteed. However, the virtual database approach is more appropriate, when the number of sources is large, the data are changing frequently, and there is little control of data sources. For these reasons, most of the recent research focuses on the virtual database approach, which is also our focus here.

When building such a web integration system, the following issues arise:

- Data modeling: an information integration system works on the pre-existing data. Thus, the first thing that the designer should do is to develop the mediated or global schema that describes the selected data sources and reveals the data aspects that might be interesting to the users. Along with global schema, the system needs the source descriptions. The descriptions specify the mapping of the global schema with local schemas at the data sources. The source descriptions served as arbitrators in case of contradictory, overlapping, semantic mismatch, and different naming conventions where different names refer to the same object. Thus, the system needs expressive and flexible mechanisms to describe the data. There are several techniques proposed for those purposes. The most well known technique is using XML with a shared DTD. Others work with ontology, for instance in the RDF (Resource Description Framework) method. Some others work with a knowledge representation language based on mathematical logic.
- Query reformulation, optimization and execution: because the user poses queries to the global schema, an information integration system must reformulate the user queries into the sources queries. Clearly, as the language for describing the data sources becomes more expressive, the reformulation process becomes more difficult. Furthermore, the system needs a query execution plan, which is the task of a query optimizer. The query execution plan specifies the order for performing the different operations in the query, and the selection of the algorithm to use with each operation. The task of the query optimization engine is difficult, for the following reasons. Firstly, the quantity of data on the Web is huge and autonomous, which means that the statistics about the sources (whether they are reliable or not) are not known in advance. Secondly, the structure of the data varies greatly, ranging from

semi-structured to unstructured data. and sources strongly differ in their processing ability. Thirdly, the data on the Web and their structure are constantly changing. Fourthly, the time needed to access the data may vary across the sources and time. After the query execution plan is completed, it is passed to the query execution engine.

– Wrapping the data sources: a wrapper is a program that reformats the data from the sources into a format that is usable by the query processor of the system. The wrapper extracts the data into a suitable source schema. An example is when the source is an HTML document. In this case the wrapper needs to extract a pre-specified set of tuples[12] from that document. Clearly, if the data and structure of the data changes frequently, manual development of wrappers is not feasible. As we have mentioned above, the development and the widespread use of XML will help to solve this problem.

Thus, web integration systems are different from typical heterogeneous database systems. There are several ways in which data mining techniques can help to alleviate some of the problems encountered above. Currently, the web mining techniques are mostly used in developing wrappers, which automate the mapping of the sources' data to the source schemas. Recently, a data mining approach is used to learn the mapping of the global schema into the source schemas, which is the task of the query reformulation engine.

Some practical applications of web mining

Example 1: ResearchIndex

ResearchIndex[13] is a digital library for scientific literature in the computer science domain. One of the most interesting features of this site is that the construction of the library is highly automated. For instance, the system automatically searches the Web for on-line papers that are relevant to computer science, tries to extract all kinds of information on those papers (such as title, authors, sub domains of computer science for which the paper is relevant, etc.) and stores the information in a database. For this search, a mix of content mining and structure mining is used.

Perhaps the most impressive (at least from the information extraction point of view) software component underlying ResearchIndex is its Autonomous Citation Indexing technology. Citation indexing is the process of gathering and storing information about which articles are cited by which other articles. This kind of information is very valuable to researchers, for instance because it allows them to easily find related articles starting from a given one (note that the reference section in an article gives an overview of the papers cited by the article, but not of papers citing it). In a summarized form it is also useful for bibliometry, the science that studies how articles and authors refer to one another

12 Data objects (rows) containing two or more components.

13 http://citeseer.nj.nec.com

and that is sometimes used to estimate the impact of journals or publications. Citation indexing is important enough to allow for the existence of specialized companies that continuously gather such information manually and regularly publish new databases with this information. See also Section 2.2.3, Science mapping from publications.

The Autonomous Citation Indexing technology of ResearchIndex employs automated techniques to gather the necessary information. For each article in its database it tries to find the references section and to extract from this section, for each reference, the authors, title, journal where it appeared, publication date, etc. This information is again stored in the database. Under the assumption that most of the information has been extracted correctly, the database can then give relatively accurate answers to queries such as 'list all papers that contain 'data mining' in their title', 'which papers have a reference to paper X?', or 'how many articles, authored or co-authored by person Y, have been published in the year 1999?'.

Obviously, the task of automatically extracting author names, article titles, etc. from documents is difficult, and errors do occur in the database; but if one is interested in general statistical information (as for bibliometry), then good approximations of the actual values are obtained, whereas if one is interested in specific information (e.g. which articles are citing my article?) then in many cases incorrect results are still sufficiently interpretable by the user to be of use. Thus, while the information offered by ResearchIndex is not perfect, it is certainly good enough for practical purposes. Information on its Autonomous Citation Indexing technology is available at http://www.neci.nec.com/~lawrence/aci.html.

Example 2: Creating corporate profiles

This is an example of integration of information from heterogeneous and distributed data sources[14].

Finding and integrating information from different sources is time-consuming and error prone, when done manually. Consider, for instance, the task of a financial analyst. For a given company, the analyst needs to gather information related to this company from various resources. These resources can be internal data, which is corporate engagement database with that company, an EdgarScan database that contains financial performance derived from U.S. Securities and Exchange Commission (SEC) fillings that is available internally, and external data that is nearly real-time data such as on-line newspapers, financial web sites, etc.

The related company data from the above sources have to be retrieved, extracted into a corresponding source query, resolved from semantic conflicts and integrated. Web mining techniques could be used to build the information retrieval and extractor part, since building it manually is time-consuming and not scalable.

14 http://context.mit.edu/~coin/demos/

There have been some intelligent information agents and information extraction systems built to deal with the above problem. The ongoing work on a 'semantic web' will make the resolution of semantic-conflict problems more manageable. See also Section 5.6.2, Extracting knowledge from the Web.
The resulting integrated data can serve as a starting point for further analysis or as an input to decision support systems. If needed, the integrated data can be used as an input to a data mining system, which an analyst could use to discover interesting new knowledge about that company. Another approach to company rating can be found in Section 3.2.2, Visual assessment of creditworthiness of companies.

Conclusions

The World Wide Web, viewed as a huge and heterogeneous repository of information, motivates the development of new techniques for retrieving information, techniques that are much more sophisticated than the ones typically used for classical databases. There is cross-fertilization between information retrieval and extraction on the one hand, and data mining on the other hand. Both may be useful as a component of the other. The state of the art in both domains is steadily advancing, as can be seen by several impressive applications that already exist on the Web. On the assumption that the current trend continues, it is reasonable to expect that in the next decade the Web will evolve into a knowledge base, the completeness and intelligence of which will largely surpass that of any encyclopedia, newspaper or classical library, and for many domains even that of human experts.

References

- Abiteboul, S., D. Quass, J. McHugh, J. Widom, J.L. Wiener. (1997). The Lorel Query Language for Semistructured Data. International Journal on Digital Libraries **1** (1):68-88
- Abiteboul, S., P. Buneman, D. Suciu. (1999). Data on the Web: From Relations to Semi-Structured Data and XML. Morgan Kaufmann
- Allan, J., R. Papka, V. Lavrenko. (1998). On-Line New Event Detection and Tracking. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp37-45
- Berners-Lee, T. (1998). Semantic Web Road Map. Work in progress. http://www.w3.org/DesignIssues/Semantic.html
- Bex, G.J., S. Maneth, F. Neven. (2000). A Formal Model for an Expressive Fragment of XSLT. Computational Logic – CL 2000. Lecture Notes in Artificial Intelligence **1861**:1137-1151. Springer Verlag
- Blum, A., T. Mitchell. (1998). Combining Labeled and Unlabeled Data with Co-training. Proceedings of the 1998 Conference on Computational Learning Theory

- Bonifati, A., S. Ceri. (2000). Comparative Analysis of Five XML Query Languages. SIGMOD Record **29** (1):68-79
- Deutsch, A., M. Fernandez, D. Florescu, A. Levy, D. Maier, D. Suciu. (1999). Querying XML Data. Data Engineering Bulletin **22** (3):10-18
- Etzioni, O. (1996). The World Wide Web: Quagmire or Gold Mine? Communications of the ACM **39** (11):65-68
- Garofalakis, M.N., A. Gionis, R. Rastogi, S. Seshadri, K. Shim. (2000). XTRACT: A System for Extracting Document Type Descriptors from XML Documents. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, SIGMOD Record **29** (2):165-176
- Kleinberg, J.M. (1998). Authoritative Sources in a Hyperlinked Environment. Proceedings of ACM-SIAM Symposium on Discrete Algorithms. pp668-677
- Kosala, R., H. Blockeel. (2000). Web Mining Research: A Survey. SIGKDD Explorations **2** (1):1-15
- Kumar, S.R., P. Raghavan, S. Rajagopalan, A. Tomkins. (1999). Trawling the Web for Emerging Cyber-Communities. Proceedings of the Eighth World Wide Web Conference (WWW8)
- Lang, K. (1995). News Weeder: Learning to Filter Netnews. Proceedings of the 12th International Conference of Machine Learning (ICML'95)
- Lawrence, S., C.L. Giles. (1999). Accessibility of Information on the Web. Nature **400**:107-109
- Lee, D., W. Chu. (2000). Comparative Analysis of Six XML Schema Languages. SIGMOD Record **29** (3):77-86
- Mladenic, D. (1996). Personal WebWatcher: Implementation and Design. Technical Report IJS-DP-7472. http://www.cs.cmu.edu/~TextLearning/pww/
- Neven, F., T. Schwentick. (2000). Expressive and Efficient Pattern Languages for Tree-Structured Data. Proceedings 19th Symposium on Principle of Database Systems. pp145-156
- Robie, J., D. Chamberlin, D. Florescu. (2000). Quilt: an XML Query Language. http://www.almaden.ibm.com/cs/people/chamberlin/quilt_euro.html

5.6.2 Extracting knowledge from the Web

Danny Lie[15]

Introduction

The common method of retrieving knowledge from the Internet is using a regular search engine. However, as the amount of knowledge on the Web increases, this method proves to be impractical: a simple query can easily result in millions of search results containing an abundance of irrelevant documents. This article proposes a new way of finding knowledge on the Web.

After a discussion of some of the problems with traditional search engines, the next two paragraphs describe a solution to this problem. Finally, we discuss how this solution can be used in practice.

Problems with traditional search engines

In traditional search engines, the user has to translate an idea to a query. However, an idea can be quite complex and not be easily converted into a simple query. Therefore, most users use one or more keywords that directly spring to mind when formulating a query. The search engine simply returns all pages in which these keywords occur. The original idea is most probably present in this abundant search result, but how will the user ever find it? Figure 1 illustrates this problem.

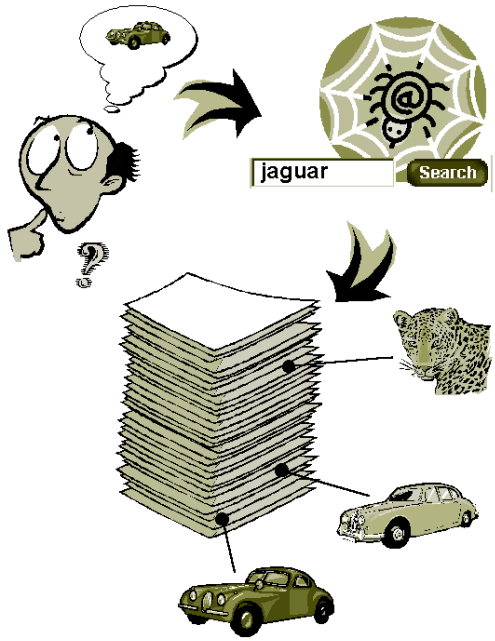


Figure 1
Looking for a needle in a haystack.

15 Ir D. Lie,
lie@carp-technologies.nl,
Carp Technologies (CEO), Enschede,
The Netherlands,
http://www.carp-technologies.nl

This problem exists because a regular search engine has no idea what the documents on the Web or the documents in the search result are about. Instead, it simply matches keywords and therefore is very limited in assisting the user in finding information. For a computer to be able to properly assist the user, it has to know what all the documents on the Web are about. This way, the computer can ask the user for clarifications ('Do you mean the jaguar as a car or as an animal?') and give suggestions ('I have found cars and animals').
However in order to understand all the documents on the Web, a computer has to understand human language.

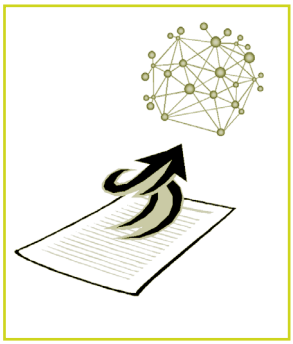


Figure 2
Extracting knowledge from texts.

Extracting knowledge from text documents

In Section 5.4.1, Understanding human language the mechanisms for extracting knowledge from texts are explained in detail. This process results in a semantic structure.
The semantic structure contains the different concepts from the text and their relationships with each other. Thus, a semantic structure represents the meaning of the text. Using the steps described in Section 5.4.1, one can automatically extract knowledge from texts as shown in Figure 2.

Building a giant knowledge network from the Web

The techniques described in Section 5.4.1 can be used to analyze every document on the Web. This results in billions of semantic structures that can be combined into one giant semantic structure. This knowledge network would contain the knowledge of the entire Web!
To build this network, one can combine nodes from different semantic structures, as illustrated in Figure 3.

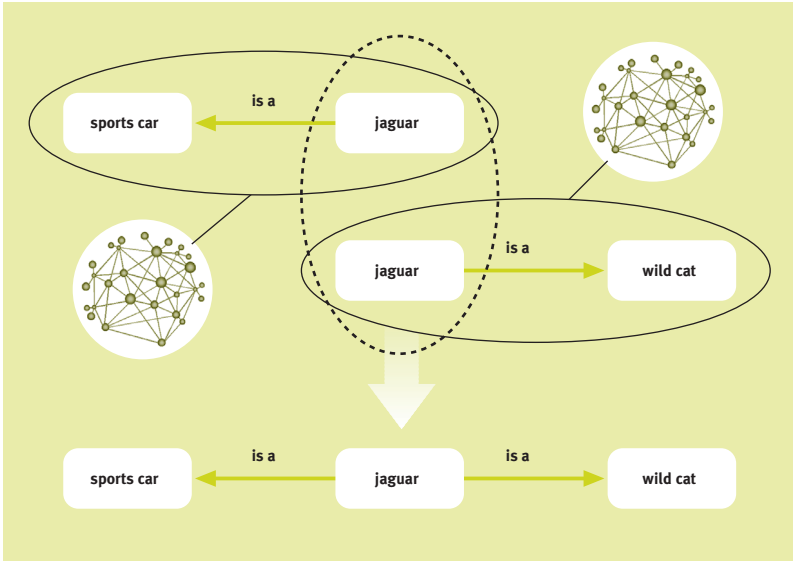


Figure 3
Combining semantic structures by combining nodes.

We can even add more knowledge by inferring relations. For example, Figure 4 shows how the knowledge 'herbivores eat vegetable components' can be inferred from the facts that 'cows eat grass' and 'horses eat apples'.

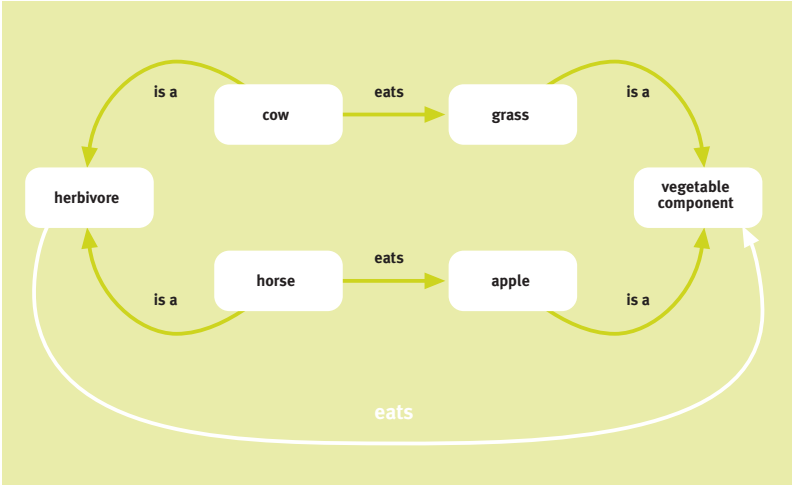


Figure 4
Inferring new knowledge from existing knowledge.

Furthermore, any erroneous information that may be contained in the individual documents that have been analyzed can be prevented from entering the giant knowledge network by using statistical filters. These filters only allow certain knowledge to enter the giant network, if this knowledge can be found in more than one document. By using these and other techniques one can automatically build a giant knowledge network from the Web as visualized in Figure 5.

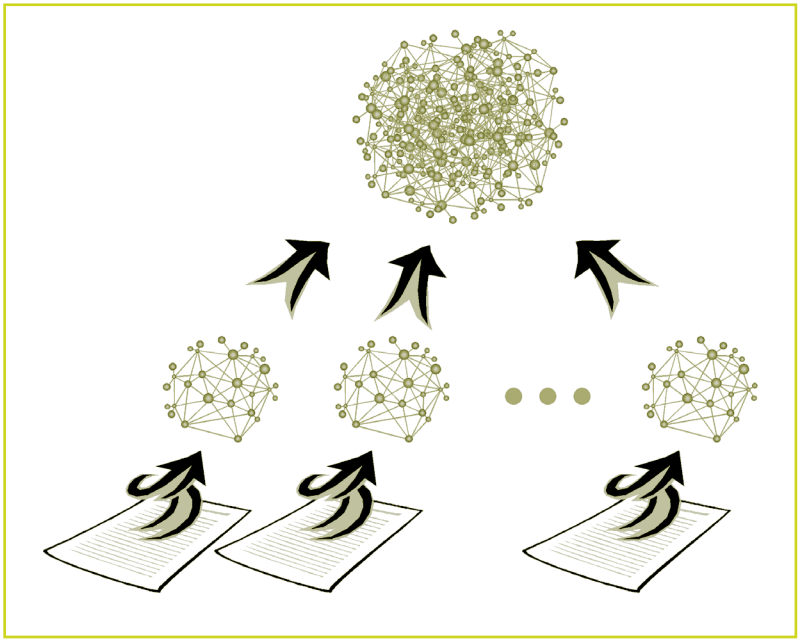


Figure 5
Building a giant knowledge network from the Web.

The resources that are required to build this knowledge network are vast, but comparable to those needed for traditional search engines. To give an estimate: we have built a prototype that processes one million documents per day, using 10 regular desktop pc's. Thus, using 1,000 pc's, one can process two billion documents in about a month.
How all this will help in building a better search engine will be explained in the next paragraph.

An architecture for a next-generation search engine

Using the knowledge from the automatically generated knowledge network one can cluster the search results into meaningful groups. For example, the knowledge network will contain the knowledge that a jaguar can be both a sports car and a wild cat, as shown in Figure 6.

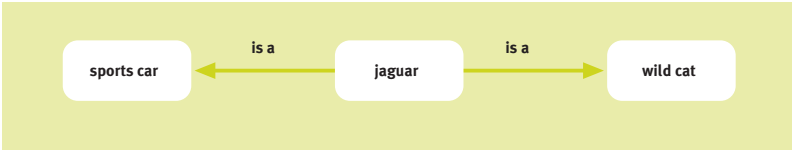


Figure 6
Example excerpt from the giant knowledge network.

This information enables the search engine to present the search results in groups of documents. When a particular group is selected, it can be iteratively clustered again into groups, which enables the user to interactively filter the initial abundance of search results to a much smaller and more relevant set of documents with a couple of mouse clicks. Figure 7 visualizes this.

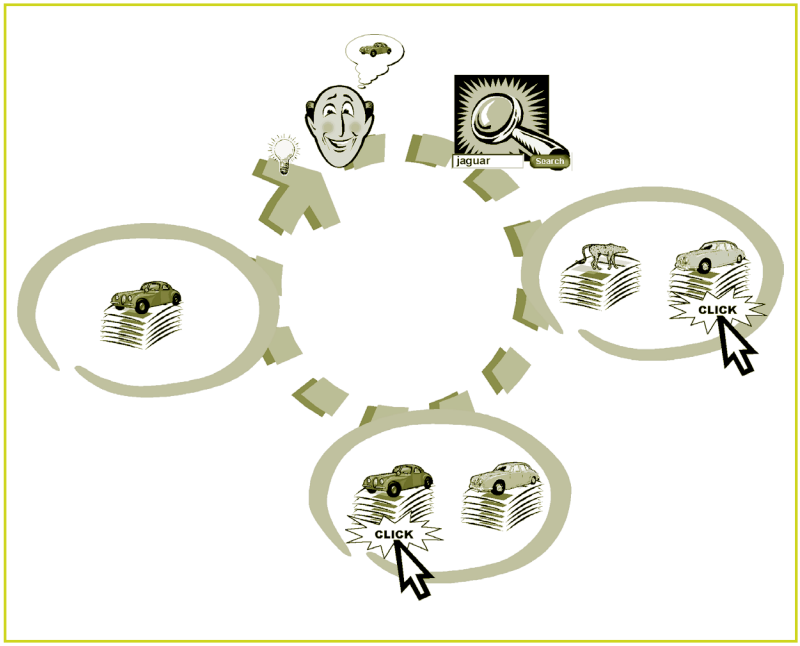


Figure 7
Finding information with an intelligent search engine.

Conclusion

The techniques described in this section can be used to automatically build a giant knowledge network by analyzing a very large number of documents. In turn, this knowledge network enables a search engine to cluster the search results into meaningful categories, which assists the user in finding the relevant information. Furthermore, this knowledge network and the techniques to generate it, can of course be used for a myriad of other applications like document classification, topic determination, document comparison, etc.

References

- Allen, J. (1995). Natural Language Understanding. Second Edition. The Benjamin/Cummings Publishing Company, California
- Carp Technologies homepage. http://www.carp-technologies.nl
- Charniak, E., Y. Wilks. (ed.). (1976). Computational Semantics. North-Holland Publishing Company, Amsterdam. pp101-184
- Reiter, E., R. Dale. (1997). Building Applied Natural Language Generation Systems. Natural Language Engineering **3** (1):57-85. Cambridge University Press, United Kingdom
- Schank, R.C. (1984). Conceptual Information Processing, Elsevier Science Publishers, Amsterdam
- Srinivasan, V. (1993). Punctuation and Parsing of Real-World Texts. Twente Workshop on Language Technology 6 – Parsing Natural Language. University of Twente, The Netherlands

5.6.3 Mining for adaptive web sites

Geert-Jan Houben[16]

In this part of the book, centered around the individual user, we now look at a specific application area of data mining: adaptive hypermedia systems. We explain what adaptive hypermedia systems are, and argue that their successful application can benefit from data mining. Specifically, with the use of adaptation in systems implemented on web sites, data mining will allow the designers of these systems to meet the demands of the future.

Personalization

As stated in [Mulvenna, 2000] the basic goal of personalization systems is to provide users with what they want or need without requiring them to ask for it explicitly. They define personalization to be the provision to the individual of tailored products, services or information relating to products or services. They describe it as a broad area, covering recommender systems, customization, and adaptive web sites.

An important motive for personalization is that some systems are best used, if the user is not required to spend a lot of effort in making the system suited to his own purposes. For example, the system of a mail order company might want to please clients by offering the 'right' products. Subsequently, the company might decide that the user should not be required to do the company's work of pleasing its clients: the users should be able to get access to the right products without being bothered by the work that the company is putting into personalization. For an example, see [Ardissono, 2000]).

[Perkowitz and Etzioni, 2000] state three main reasons for personalization and adaptation:

- First, there is the fact that users have different views of the information they are looking at in an application.
- Moreover, the perceptions of these users change over time.
- While it is already difficult for a designer to decide what the perceptions of the users will be, there is an additional difficulty in the fact that the designer's initial expectations can be violated by the actual perceptions of the users.

If we are considering an application realized as a company web site, then it is quite essential that this web site pleases everybody. It should not be just aimed at a specific group of people unless of course that is explicitly the target. Too often we encounter a 'one size fits all' approach. It is essential that every possible user is managed in a fair way. The answer to this problem is often found in adaptation as a way to realize personalization.

16 Dr G.J. Houben, g.j.houben@tue.nl, Eindhoven University of Technology, Department of Mathematics and Computing Science, Information Systems, Eindhoven, The Netherlands

In order to adapt an application to the current needs of the user, one cannot be satisfied with an approach in which a human designer manually adapts the information. Necessarily, this process needs to be automated, allowing an application to adapt without human intervention. In our discussion of adaptive hypermedia systems we therefore concentrate on the automated process of adaptation. A central question remains how to divide the labor between the 'automated assistant', i.e. the automated process incorporated in the adaptive system, and the human designer who is specifying and controlling this process of adaptation. Of course, there is also the question whether and how non-trivial adaptations can be automated.
In the rest of this section we concentrate on adaptation in the context of adaptive web sites and adaptive hypermedia applications, but we would point out that this is just one example of personalization.

Adaptive web sites

While web sites may be the better-known examples, most of the technologies described in this section refer to the more general notion of hypermedia (or hypertext). Hypermedia generally offers functionality that in current web implementations might not be possible, but in order to obtain a good overview of personalization a focus on adaptive hypermedia is essential.
If we focus on web sites and how they can adapt, we should look at the relevant aspects of that adaptation. In order to achieve adaptation, a designer can choose to base the adaptation on the contents of the application or on people's navigational choices. This major difference relates to the focus: does the designer start thinking about the adaptation from the point of view of the content (and how this content can be presented), or do the access behaviors of the users serve as the starting point for the adaptation design.

In a web site aspects of the contents that can adapt are the content that is provided on the pages, the layout of the pages that are offered, and the structure of the site. We will come back to this later. First, we pay attention to the navigation behavior. If the navigation paths serve as a basic part of the adaptation process, it is often the case that the adaptation process includes some mechanism to perform path prediction: "given the situation known, where do we expect that the user will go or want to go: let's adapt accordingly." However, in order to predict the future user behavior, one first has to 'know' the present user behavior. In that context it is worth realizing that besides the patterns of usage (in terms of navigation), other observable characteristics of the user might be taken into account, e.g. habits or preferences. We should realize that in the different applications of adaptation several different approaches could be used:

– An approach that one typically finds in portal web sites is targeted at the masses, trying to make the best out of the 'mass production'.

– One approach asks the collaborative cooperation of the users, realizing an adaptation based on explicit feedback from the users. We will not go into more detail here, but this approach appears to show interesting results (based on the idea that if you ask for a little cooperation of the user community, you can help a lot).
– An alternative approach is based on the idea that users follow the paths of previous users: "seeing that a certain path has been followed quite a number of times, attracts new visitors."

In order to perform an effective adaptation, specifically in the case of web sites, some approaches use meta-information. Primarily, they assume meta-information on the user, but also on the contents of the web site. While such meta-information can benefit a detailed adaptation (of course, when relevant information is available, it is easy to exploit that information for the purpose of adaptation), it is important to realize that this meta-information will not always be (directly) available. Moreover, demanding that this meta-information is made available is not always realistic. One of the cornerstones of adaptation is that one wants to have a practical solution to the problem of personalization. Therefore, we usually have to deal with the data that is already available.

Crucial to an adequate process of adaptation is an application architecture that allows to separate and control the relevant data in the adaptation process. For example, the contents of the application, a major factor in this process, should be dealt with in such a way that the adaptation process can be designed according to need. In order to organize this domain to which the objects of interest belong, it might be useful to use a 'concept hierarchy'. The AHAM model of [Bra, 1999b] is a reference model for adaptive hypermedia that distinguishes a domain model, a user model and an adaptation model (or teaching model, for the area of educational applications). While most of the early applications and systems had an 'integrated' adaptation process, in which these three models were not separated, their research shows the advantages of separating these three aspects for the benefit of a controllable process of adaptation. Choosing an adaptive system that allows for applications that elegantly separate these different aspects makes designing the application much easier and more effective.

In relation to this, it is important to realize that the designer of an adaptive application should still be in control (of the adaptation process). On the one hand the provider of the information, for reasons of efficiency, will try to offer a standard way of access to that information. On the other hand, the users, with their differences in usage, should be treated as if they were the only party the application is communicating with. It is this balance that a self-adapting appli-

cation seeks to fulfill. In realizing adaptation it is worth recognizing the importance of the degree to which the adaptive system can perform the adaptation automatically. The system should be able to help the human designer in automating some of the more trivial tasks in this process of personalization, while the designer should be able to effectively specify how the system reacts to the use by the individual user.

By now we have pointed out how an important part of the execution of the adaptation process needs to be automated. However, in order to be able to adapt, a system first has to know the user and his behavior. This implies a thorough process of analysis. We now discuss this analysis phase and conclude that mining can play an important role there.

Mining for adaptation

In the analysis of user-behavior mining can play an effective role: it is however only a first step. After the mining the results need to be represented and then deployed.

Before we pay more attention to the analysis phase, first an observation concerning the representation phase. More and more applications and systems use an XML-related format in the representation process. This means that the data is made available for the actual deployment. One advantage of this movement toward this XML-based representation is that along with it goes a standardization of the data to be used for the deployment. In the spirit of the standard and reference models, like AHAM, a uniform mechanism exists to represent the data to be used in the adaptation process. In the area of educational applications this allows, for example, the exchange of user models between different applications, i.e. different courses. Specially, in the context of the Web, where adaptation can offer new possibilities, the advantage of standardization is obvious. As a consequence the 'new' medium opens up more and new applications of adaptation.

The analysis-phase targets the collection of data relevant for the adaptation process. The most essential part of knowing what the users want is to obtain an insight in what they are doing, i.e. what they have been doing in the past. Knowing what a specific user has done can give a hint as to what the next steps of that user might be. Knowing what all users have done in the past can give an insight into the entire and global use of the application or site. On the basis of these insights, fed by the intelligence demonstrated by the human users, it is then possible to exploit this and have the application automatically select the right information to proceed with.
It is a fact that this often leads to a tremendous amount of (web usage) data

available for the process of (specifying) adaptation. Specially, in the context of customer relationship management or marketing purposes, the target for the future is to come up with practical solutions to derive relevant data from the vast amounts of data collected. As an indication, as is for example illustrated by [Spiliopoulou, 2000], the preparation of the web log for analysis (even before the actual mining) requires extensive preprocessing.

The data available as a basis for the adaptation is usually the data that describes the user-access patterns from the past. This data includes information on which pages have been viewed, which links have been followed (by registering the 'clicks'), and the identification of sessions, i.e. coherent access in the context of one specific task or goal. Usually, this data is not publicly available. However, most systems log these data and subsequently they are available for the application designer to consider improvements to the application. Most practical systems base their adaptation on these usage logs or traces. One major disadvantage in this context is the fact that usually the designer is not able to determine who is actually the user. Sometimes there is something like an IP-address available or a cookie, but there is no absolute guarantee as to who is in fact the physical person behind this 'virtual address'. Often, the designer is not so much interested in determining the actual user per se, but what a designer does want to know is whether some interesting or unexpected patterns of user behavior can be detected: they could tell whether some design action is necessary or advised. These design decisions will then seek to improve the validity of (pedagogical) decisions implicit in the application's design and content. Section 5.6.1, An overview of web mining gives more insight into the use of mining in the context of web applications.

Adaptive hypermedia

Adaptive hypermedia is a new direction of research within the area of user-adaptive systems. The main goal is to increase the functionality of hypermedia by making it personalized. Adaptive hypermedia systems build a model of goals, preferences and knowledge of the individual user and use this throughout the interaction to adapt to the need of the user (and the application). This is specifically relevant for hypermedia applications that are to be used by people with different goals and knowledge and where the hyperspace is reasonably big.

As a working definition we refer to adaptive hypermedia systems as systems that reflect some features of the user in the user model and apply this model to adapt various visible aspects of the system to the user. In this section we use parts of [Brusilovsky, 1996] (see [Bra, 1999a] for a nice summary), which is an excellent, comprehensive review of adaptive hypermedia systems. For those

readers that are interested in actual implementations of adaptive hypermedia we refer to that review. Here, we restrict ourselves to a high-level overview of the definitions and concepts used in that review, since they help us describe this area of adaptive applications and the connection to mining. Note that in this section we will pay more attention to the process of adaptation than to the process of user modeling needed to perform high-quality adaptation. A lot of research is generally devoted to user modeling (and it is in that aspect that mining plays its primary role), but we want to concentrate here on the specific aspect of applying that user knowledge in adaptation.

A characteristic target is to obtain a system that allows for an automatic or semi-automatic process of improving the organization (structuring) and presentation of information in the application or site. A nice example is the construction and maintenance of index pages for an application: to make that a process that continuously adapts to the current situation asks for a structured approach. In a similar context we can make the remark about the difference between customization and transformation. By this we mean that, as opposed to actually transforming the application, customization refers to 'non-destructive transformations': they try to change the perception of the user without really reconstructing the information.

In this section we address several different kinds of methods and techniques used in adaptive hypermedia. As pointed out, [Brusilovsky, 1996] gives an excellent introduction in the field, and we use his definitions here to explain the different kinds of adaptation possible. While the difference between adaptation techniques and methods is not crystal clear, we can use the following working definitions. The techniques of adaptation are part of the implementation level of an adaptive hypermedia system. They are characterized by a specific kind of knowledge representation and a specific adaptation algorithm. Usually, these techniques are closely related to the systems in which the applications are operating. On the other hand, the methods of adaptation are generalizations of adaptation techniques. Such a method is based on a clear conceptual idea of adaptation. Together, the techniques and methods build the entire collection of tools that an application designer can use to make an adaptive application.

We warn here that adaptation relies heavily on the concept of relevance, and that is very subjective. It is quite easy for a designer to introduce a potential mismatch, which will have a direct impact on the effectiveness of the application.

To qualify and characterize the different aspects of adaptation it may be useful to ask the following questions:

– Where can the adaptive systems be helpful?
– What features of the user are used as a source of the adaptation?
– What can be adapted, related to the technologies of adaptation?
– What are the goals of the adaptation, related to the application area?

Application areas

In order to answer the question where adaptive hypermedia systems can be useful, [Brusilovsky, 1996] identifies a number of application areas, some of which we will discuss here.
One of the most popular areas, and also one that can illustrate the usefulness quite well, is educational hypermedia. Typically, these applications concern a small hyperspace for a 'closed' domain, in which the designer/teacher tries to educate the users/students by providing them with contents depending on their knowledge of the domain. The teacher can use adaptation to respond differently to students with a different background or state of knowledge. Moreover, the adaptation can be used to offer the students help in navigating through the hyperspace. Well-known examples of educational applications are courses on hypermedia or on graphical user interfaces, offered by Eindhoven University of Technology, based on the adaptive hypermedia system AHA [Bra, 1998].
On-line information systems form another illustrative example area. Their aim is to provide reference access to information to users with different degrees of knowledge about a subject. Here, the hyperspace is not so much inspired by the educational process as in the case of the educational applications, but the contents in the hyperspace collectively try to offer all the information that the users might want to query. However, the state of knowledge of the users might again be very different, and that is exactly the reason why the application designer might want to differentiate between the different users.

Other classes of adaptive systems that [Brusilovsky, 1996] distinguishes are on-line help systems, information retrieval hypermedia systems, institutional information systems, and systems for managing personalized views in information spaces. Note that in itself any hypermedia application is already an adaptive system: it allows the user to browse through a hyperspace, and by navigating through that hyperspace the system can react by offering specific contents.

Features to adapt to

Knowing the application areas, a designer could derive features of those users to which the application or system adapts. These features describe to the designer, which user the system is communicating with.

The aspect that comes immediately to mind is the knowledge the user has on

the contents or subject of the application. In the case of an educational application [Bra, 1998], it is straightforward that the system as a representative for the teacher might want to assess what the student knows before offering some (new) material or contents. In order to be able to manage the adaptation on the basis of this knowledge, the designer must instruct the system how to build up and maintain a model of the user's knowledge. Several approaches exist to model the user's knowledge: overlay modeling and stereotype modeling are just two of these approaches. All of these modeling approaches try to associate an assumed level of knowledge to the concepts belonging to the domain. They differentiate in the way in which they categorize the concepts from that domain and in the level of detail in which they express the assumed level of knowledge. Another feature that can play a role in the adaptation is the task the user is performing. The different application areas have different ways in which they implement the task-based approach, but it is obvious that a system that knows (or thinks to know) what a user is trying to achieve, can improve its services (by adapting). Often, the task at hand is related to a goal, and therefore this kind of adaptation is also known as goal-based adaptation. In the context of an educational application setting goals (and adjusting the adaptation to it) can help the teacher in controlling the educational process concerning the specific student. Just like knowing the user's task or goal, knowing the user's background and experience can also play a beneficial role. Background refers to the knowledge and information outside of the actual subject of the application, but which is relevant enough to be taken into account. Experience describes how well the user knows the hyperspace and how comfortable he is in navigating it. Again, this requires an application designer to assume these characteristics per user and then to adjust the adaptation accordingly.

User preferences are the last type of feature that [Brusilovsky, 1996] mentions. It is obvious that the user can have preferences as to almost all of the aspects of an adaptive hypermedia application. Specifically, these preferences concern the contents themselves, as well as the way in which the system is performing this process of offering contents. Since user preferences cannot actually be deduced by the system, the system should take precautions to get the information necessary to accommodate the preferences: usually, this leads to some dialogue in which the user identifies to the system what the preferences are. As a consequence, one could see preferences more as an example of adaptability than adaptivity. By this difference we mean that adaptability refers to a situation in which the user is able to specify his influence ('in one step') and then the system is adjusted to that, while in the case of adaptivity this process of (registering) user influence and then adapting to it continues during the user's usage of the system.

Adaptation aspects

To answer the question what aspects can be adapted, it is fairly obvious that in the context of hypermedia applications two aspects are prominent candidates. Since a hypermedia application or hyper-document consists of pages representing contents and links allowing the user to navigate between these pages, it is clear that adaptation could be realized both at the content-level and at the link-level.

Adaptive presentation aims at adapting the contents of pages, based on the user features that the designer wants to take into account. A straightforward example is that a novice (to the domain) asking for the contents of a subject is presented with contents that explain all the relevant details, while an experienced user might just get the core information based on his experience. In 'classical' hypertext applications the changes in the contents imply changes in the text that is offered: most adaptive presentation is implemented by adapting the text that is visible at a certain page to the situation at hand. However, in hypermedia (or multimedia) applications the changes can also affect the multimedia items that are shown. [Rutledge, 2000] addresses exactly this issue of adapting the multimedia content.

While adaptive presentation received a lot of attention in the early days of hypermedia, the next phase in adaptation seeks to exploit the typical 'hyper-aspects' of hypermedia: by adapting the navigation support it is possible to help the user in better navigating through the hyperspace. [Brusilovsky, 1996] reviews several technologies for this purpose. Since this navigation support depends heavily on the presentation of hyperlinks, it is wise to start by considering different ways of presenting links:

- The typical link on a regular page is a local non-contextual link, which means that the link (or more precisely, its source anchor) is visible on the page independent from the content. Such links can easily be changed in the light of adaptation, by sorting, hiding or annotating them.
- Contextual links have source anchors that are so called 'hotwords' (in text) or 'hot spots' (in pictures): these anchors cannot be easily changed, except in terms of their annotation.
- Index pages and content pages are special pages that only contain links (for the sake of offering a rich list of links to the user). Usually, these pages have a specific order for these links, which limits the possibility for adaptation.
- Local maps and global maps usually represent the structure (of the local environment or the entire hyperspace) in a graphical way, such that the user can use this map for direct navigation.

The most trivial technology of adaptive navigation support or link adaptation is direct guidance. That means that the system tells the user what the next page should be. By offering a link to this 'next page', the system can guide the user in his navigation. While this approach is easy to implement, it offers only limited support.
Adaptive ordering sorts all links on a page in such a way that the most relevant is in the most prominent spot on the page. While this is excellent for some purposes, one of the known disadvantages is that the ordering can change over time and thus the (novice) user might get a very different view of the page each visit and thus get confused. Several researchers have investigated this phenomenon, specifically the question in what circumstances this adaptive ordering is effective, e.g. in information retrieval applications or on-line documentation systems.
The most often used technology is hiding: showing only the relevant pages and not showing (and thus not giving access to) non-relevant pages is one way of helping the user navigate. It appears that hiding is easier for the user (when compared to sorting), as the adaptation usually implies incremental changes to the link presentation.
By adding a comment to a link, telling the user what to expect at the end of it, we can annotate links. Adaptive annotation tries to help the user by manipulating these comments. Note that these comments do not necessarily have to be textual, but they can for example be visual, e.g. different icons, colors or font sizes. Since annotations are an easy way to indicate properties of a link (or destination node), a lot of research has tried to investigate the effect of certain annotation approaches.
Finally, local and global maps can be used as a way to adaptively communicate with the user: this is called map adaptation. This basically implies changes to the (graphical) appearance of the map.

Next, we discuss methods and techniques that help to provide content adaptation and adaptation to navigation support.

Content adaptation methods

The method of additional explanations offers the application designer the possibility of providing a user with an additional piece of content, when this is appropriate. A typical example would be some piece of information that would be irrelevant (and even confusing) for a novice user: by hiding it for this novice and showing it to the experienced user, the system can make sure that the display of this piece of information serves its purpose.
Similarly, prerequisite explanations and comparative explanations offer explanatory pieces of information to a page when this is relevant. In case of a prerequisite explanation the explanation is added, when the user 'needs' this

explanation in order to understand the rest of the page: the explanation tries to overcome an apparent deficiency in knowledge. Comparative explanations are added when a comparison makes sense, since the user visited a comparable concept earlier.
When the previous adding and deleting of pieces of a page is not sufficient, explanation variants may offer the freedom to choose between different variants of some of the page's elements.

Content adaptation techniques

In order to realize the adaptation methods mentioned several techniques for content adaptation exist. The most straightforward one is that of conditional text. The different pieces of the contents are made conditional by specifying some condition stating, when the piece is to be displayed or not. While this technique requires a rather detailed view of the different pieces of the information, it turns out to be quite flexible and effective for the designer. For every small piece of contents, the designer is able to specify under which circumstances this is shown.

Stretchtext is an approach that differs from the 'traditional' approach known from the Web, since it reacts differently to a user's click on a link's source anchor. In the average web application the system will respond by replacing the current page by the one that is the destination anchor of the chosen link: this implies that the original page disappears (or is visible in a separate window). In stretchtext the source anchor is replaced by the content (text) that is associated with the destination anchor, but the contents surrounding the source anchor remains 'in place': this implies that the text of the original page is extended at the place of the activated link with the text associated with the destination of that link. For the sake of adaptation the information is presented first with all the possible stretchtext-extensions 'closed' and gradually they are opened up, when appropriate. This approach also allows for closing such extensions whenever they are not relevant anymore.

Page variants and fragment variants are techniques that allow for the maintenance of different variants at the level of page and fragment of a page respectively. It is obvious that fragment variants are a more refined method, but the consequence of this flexibility for the designer is the need to carefully compose a page out of fragments.
The frame-based technique allows for the manipulation of the contents in the form of frames. The contents of a frame can then be expressed by the specification of what happens to the different slots of the frame: which slots are presented (and in which order) and what do they contain.

Navigation support adaptation methods

Global guidance refers to the support that the system offers the user in finding the way to the information necessary for reaching the global goal. Knowing the global objective of the user, the system can help in offering the shortest way to get there, without unnecessary side steps. Global guidance can be achieved by telling the user at every single step where to go, but this guidance can also be achieved at a conceptually higher level. In an educational application this might lead to a schema in which the student is guided to the learning goal in the fastest possible manner. In a situation where the user would be looking for information this guidance should be dealt with quite differently.

As opposed to the higher-level objective of global guidance, local guidance tries to help the user in determining the very next step. So, the objective is much more dependent on the local environment (not on the overall objective of using this application).

Local orientation support is a method that aims at providing the user with a sense of orientation: while guidance tells the user what to do next, orientation support tells the user where he is and what is around him. Local orientation support informs the user about the pages reachable from the current one. In an educational application the system might use a technique like hiding to offer the user a view in which only the relevant pages are accessible: here, relevance could mean that the page is 'ready to learn'. The target is that the user is able to make an appropriate navigation choice.

On the other hand, global orientation support is used to assist the understanding of the overall perception of the hyperspace. So, instead of a local view on the immediate environment of the current page, global orientation support involves techniques that help to represent the global 'map': e.g. by actually providing graphical maps, or by introducing visual aids such as landmarks or sign posts.

Adaptive navigation support techniques

Sorting links is one of the most used techniques to adaptively support navigation. By sorting the available links on the basis of their relevance, i.e. the relevance of their destinations, the methods of guiding the user towards the next page can be realized.

Hiding links and annotating links are other techniques that can be exploited to support guidance and orientation support. In educational applications it can be helpful if the system informs the user about which pages (or concepts) are 'already learned', 'ready to be learned', 'not ready to be learned', etc. Several solutions exist to actually inform the user about this: e.g. using red and green coloring of the links to specify the way to go.

Conclusion

After considering the diversity of methods and techniques for adaptation we return to the crucial role that user modeling plays in this process. In order to be able to have an adequate process of adaptation it is essential that the right data is available to base the adaptation on. The analysis of previous usage data is one of the ways to achieve this. These usage data, as stated earlier, appear in a format that contains a lot of detail, but that does not offer a high-level view of the user behavior.

While some advantage can be obtained by some kind of collaboration of the users, either individually or collectively, that approach does not suit the needs of all applications. Some applications cannot place this burden on their users. Therefore, in these applications, usage data is the source of information that should be exploited.

It is exactly for this reason that data mining can play a role in adaptive hypermedia. The future developments of adaptation can therefore benefit from further developments in data mining.

References

- Ardissono, L., A. Goy. (2000). Dynamic Generation of Adaptive Web Catalogs. Adaptive Hypermedia and Adaptive Web-Based Systems. Springer LNCS 1892. pp5-16
- Bra, P. de, L. Calvi. (1998). AHA! An Open Adaptive Hypermedia Architecture. The New Review of Hypermedia and Multimedia. pp115-139
- Bra, P. de, P. Brusilovsky, G.J. Houben. (1999a). Adaptive Hypermedia: from Systems to Framework. ACM Computing Surveys **31** (4es)
- Bra, P. de, G.J. Houben, H. Wu. (1999b). AHAM: A Dexter-Based Reference Model for Adaptive Hypermedia. Hypertext'99. 10th ACM Conference on Hypertext and Hypermedia. pp147-156
- Brusilovksy, P. (1996). Methods and Techniques of Adaptive Hypermedia. User Modeling and User-Adapted Interaction. Kluwer Academic Publishers **6**:87-129
- Mulvenna, M., S. Anand, A. Buchner. (2000). Personalization on the Net Using Web Mining. Communications of the ACM **43** (8):123-125
- Perkowitz, M., O. Etzioni. (2000). Adaptive Web Sites. Communications of the ACM **43** (8):152-158
- Rutledge, L., B. Bailey, J. van Ossenbruggen, L. Hardman, J. Geurts. (2000). Generating Presentation Constraints from Rhetorical Structure. Proceedings ACM Conference on Hypertext and Hypermedia. pp19-28
- Spiliopoulou, M. (2000). Web Usage Mining for Web Site Evaluation. Communications of the ACM **43** (8):127-134

5

5.7 Mining and Personal Knowledge Management

P.P.J. Ramaekers[1], *P.M. van Rosmalen*[2]

Introduction

During the last decade the amount of accessible information has grown at an incredible speed. It is often the case nowadays that either the content of a job changes regularly, or people change jobs regularly. This causes a continuous challenge in selecting the appropriate information resources to maintain and extend people's personal knowledge.

This article explores the impact of data mining and text mining on personal knowledge management in the next decade. Starting from a description and analysis of the use of intelligent tutoring systems, current knowledge management practice and the theory of knowledge trees, as well as a theory of life long competence building from the French philosopher Serres, a synthesis is outlined which shows the challenging opportunities that (text) data mining offers to foster knowledge management and competence building from a personal perspective.

Knowledge and competence building is definitely not a linear convergent process. The importance of a holistic and cross-disciplinary view in tackling problems and challenges is well recognized. A creative attitude to innovation and problem solving is a competitive prerequisite. Therefore in the second part of this paper special emphasis is given to dealing with the use of loosely structured and tacit knowledge and the support of creative processes.

1 Dr P.P.J. Ramaekers,
info@tiaram.nl
Director Tiaram B.V., Weert,
The Netherlands,
http://www.tiaram.nl/

2 Drs P.M. van Rosmalen,
Director Quality and Innovation,
Aurus Knowledge & Training
Systems BV, Maastricht,
The Netherlands,
http://www.aurus.nl

Intelligent tutoring systems

Three decades ago, in the early seventies, the use of computers to capture and transfer knowledge began. From a relatively small but influential research area, artificial intelligence, the first knowledge based tutoring applications appeared. In contrast to the first generation of computer assisted instruction programs, which offered simple automated instruction, intelligent tutoring systems [Wenger, 1987] used artificial intelligence approaches to capture and deal with aspects of knowledge. Microworlds were shaped. Built in various ways but in general containing at least a detailed domain or expert model, a personal or student model and a knowledge transfer or instructional model. Persons involved in such a microworld can acquire new knowledge actively or in a guided way. They can immerse themselves in a device simulation or a programming world and practice their skills, as well as receive feedback and guidance depending on their progress. Alternatively, they can be guided through the study domain, while the best fitting 'chunks of information' are presented (according to their knowledge level and the instructional methods applied). The intelligent tutoring systems that have been built to date are qualitatively strong and prove to be highly effective. However, what they offer are small chunks of information and knowledge from small-scale worlds and thus have limited applicability on a real-world scale. Also, because in general they are all built from scratch, little or no effort is being paid to productivity concerns. Starting in the beginning of the nineties, steps were made to design and develop authoring systems for intelligent tutoring systems [Murray, 1999] and to deal with generic approaches, e.g. how to use task and domain ontology [Mizoguchi, 1996] to support reusable components and how to use agent architectures, which enable agents (e.g. a learner modeling agent [Paiva, 1996]) to be reused in different settings and by several types of applications.

With the emergence of the Web, research also moved to adaptive and intelligent technologies for web-based education [Brusilovsky, 1999]. Part of this research focuses on problem solving, i.e. exploring microworlds. However, the most applied examples —also the most important ones for our case — are hyperspaces of educational material. The goal here is to guide the students through the material and show them the optimal path or the optimal content. This can be achieved in different ways. The most popular use direct guidance, i.e. they offer the best page given the student's current knowledge and learning goal. This is done through adaptive link annotation and hiding (i.e. annotating the most suitable links and disabling a link, if a page is not yet ready to be learned). Additionally, based on the characteristics of the Web, student modeling[3] is extended with the option of comparing the models of different students. This gives two benefits: it becomes possible to find competent peers to share and

3 The objective of student modeling is to present the right information at the right time. Looking into more detail a classification of functions can be made, e.g. corrective functions to correct or elaborative functions to complete the knowledge on a topic. Depending on the program the functions rely on different techniques ([cf. Lehn, 1988] for a description of techniques applied in the eighties) by applying a mixture of information on the student, the domain to be covered, the instructional knowledge and an expert (the ideal student).

discuss issues. Also students who progress through their learning differently from their peer group members can be identified .

Knowledge management

In the nineties the availability and the dependence of information and communication technology drastically increased, the economy globalized and in line with these developments the importance of knowledge management was 'reinvented'. People, their knowledge and skills are again the main capital of a company. Just as in the Middle Ages, in the master model, it is the knowledge and skills of the people — the masters and apprentices — that are acknowledged to contribute most to a successful future. This started in the early nineties. First focusing on 'the learning organization', i.e. an organization that qualifies by taking learning and experience into the daily practice, later developing to include knowledge management as a corporate goal.

The importance of preserving and disclosing corporate memory in order to assist survival in an increasingly competitive environment became widely recognized, and such methods as workflow, document management systems and group-ware tools have been introduced to capture and exchange knowledge. This has been supplemented with practical techniques for building knowledge management systems that focus on both organizational and technical concerns [Tiwana, 2000].

However, whereas the objective of an intelligent tutoring system is concerned with learning and the learner only becomes actively engaged at the time of learning, knowledge management is different. In contrast to the microworlds the information preserved and disclosed is generally abundant and not very well tailored to the user. The users themselves have to decide between relevant and non-relevant information and are also responsible for maintaining and filling the knowledge management system. In both activities, finding relevant information in a growing knowledge base, and capturing knowledge (that otherwise would remain tacit) in order to make this available to others, is a tedious job.

Different solutions to these problems have been implemented or are being developed [Bair, 1998; Merlyn, 1998]:

- improving the structuring and clustering of the information;
- improving information retrieval by offering one integrated architecture to a number of independent resources [Schmitz, in preparation];
- automatically adjusting to a (pre-defined) user profile;
- automatically searching for relevant information, to a more intuitive graphical user interface (Figure 1).

For each type of interaction, i.e. gathering, organizing, elaborating and communicating information, there are tools available to facilitate the process. However, it must be realized that they have to work in the complexity of organizations somewhere balancing between order and chaos [Syed, 1998]. So it is crucial to employ technologies that minimize the time to be used by the knowledge worker. At this point two areas of tools and research are of particular importance. These are research and tools for personalization and for automatically capturing information to build knowledge bases.

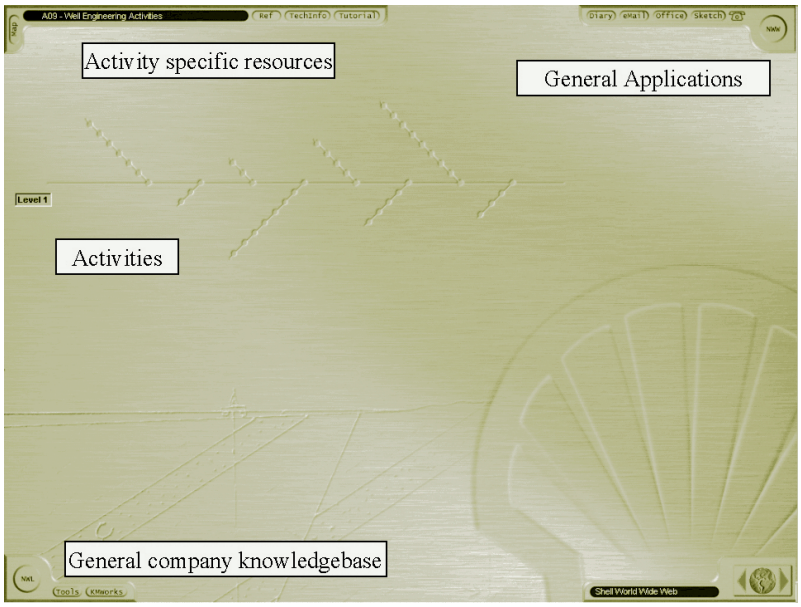


Figure 1
A snapshot of a demonstrator developed to illustrate the benefits of an integrated access to all resources required for carrying out a selected activity. Each node of the fishbone is an activity. Upon selection of a node, either the activity is further broken down or all relevant resources are activated. The resources may include handbooks, software, discussion groups, FAQ, best practice, research reports and available experts and peers.

Personalization

Probably the best-known example of personalization is the Amazon bookshop[4] [Albert, 2000]. It is based on a data mining technique called nearest neighborhood or affinity grouping or clustering. Once customers are registered, a profile is kept of their interests and books ordered. The profiles are compared and clustered. The purpose of this is to give an individual advice to each customer, i.e. an advice to have a look at books that have been ordered by people with similar interests. This approach uses little knowledge about the topic involved; it merely concentrates on similarities in interest. As a result there are some potential drawbacks. New information items can only be recommended after they have been 'recognized' by a sufficient number of users before the nearest neighborhood method can start to work. Moreover, it will only work, if you are a user whose profile has sufficient overlap with other users. If not, you will be literally lost in your own private space.

Another approach, content-based profiling, relies on the representation of content by a suitable set of attributes. The user profile is represented in a similar

4 http://www.amazon.com

way. It builds up in line with the items a user likes. The content-based filtering methods select content items that have a high degree of similarity to the user's profile. In this case the problem is to build the representation and the fact that for first time users their profile at start will be small. Subsequently, a user will receive relatively few suggestions. In [Smith, 2000] an application, a television listing service, is described that successfully combined both approaches.

Content creation

Supporting the creation of content can be done in various ways:

- offering templates or predefined structures to add information to fully automatic content 'creation' based on a combination of searching and clustering at the level of keywords;
- automatically analyzing, clustering and summarizing e-mails or documents using text data mining tools (cf. also the Paragraph Transforming data into meaningful knowledge).

Text mining uncovers relationships in unstructured collections of text documents. Applications of text mining include clustering, visualization, information extraction and summarization. It can be applied to e.g. analyzing incoming e-mails for customer support or to analyze large quantities of documents for domain specific knowledge [Lawson, 1999; Stoner, 2000].
Here one point deserves extra emphasis. It is important to assure transparency of access. Therefore the output of the analysis should be transparent, if information is required by more persons. This can be achieved by using an ontology [Gruber, 2000; Jasper, 1999], i.e. a vocabulary of terms and their relations including a specification of their meaning.

Knowledge trees

Also in the early nineties the French philosopher Michel Serres [Pouts-Lajus, 1993] was given the responsibility of considering the conditions in which an Open University could be established in France. He proposed a system in which each individual's knowledge and know-how was represented in a set of badges, a blazon. The blazon of an individual would depend on the community to which the individual belonged. In other words, it would be relative to the role within and the level of the community. Moreover a notion of time, i.e. when in time specific knowledge was acquired, was taken into account. Here a knowledge tree is the sum of the knowledge of the individual members of a community. A community and its members can use their knowledge tree to formulate their demands in terms of their ideal profiles for a given situation. The approach clarifies why, and which, knowledge has to be transferred and enables a highly individual approach.

Triangle model — a synthesis

At first sight the progress made in the two areas discussed, i.e. intelligent tutoring systems and (group) knowledge management and the theory of knowledge trees, is interesting, but they remain independent areas focusing on various aspects of knowledge. However, in the area of personal knowledge management things are becoming intertwined. The goal of personal knowledge management is to enhance our personal and business life through life long learning. This may be achieved by formally accredited courses. However, far more important is learning by doing, learning from peers and relations, and learning by access to all kinds of digital or printed resources (Inset 1). It is exactly here where intelligent tutoring systems, knowledge management and knowledge trees interact.

Inset 1: An example of a personal knowledge agent at work in interaction with an agent for the domain model, the user model and the knowledge transfer model.
March 2012: Ann recently started her new job at the microelectronics department. For the first two months her task was to get acquainted with the company and to become aware of all the issues and knowledge involved in running the company. Her first task today was to instruct her personal knowledge agent to merge her personal knowledge profile with the company's knowledge tree and her function requirements. Next she had to negotiate the access rights to the various parts of her knowledge profile. After this she left for a break. At this time her personal knowledge agent invoked a knowledge transfer process. Her personal knowledge agent entered into a dialogue with the user model agent, the domain model agent and the knowledge transfer model agent to prepare her personal internet with items of interest she could browse through and select for study, if she so decided. When Ann returned, her computer showed a map of items of interest to study, of people she could contact and a visualization tool pad to structure her thoughts as she goes through her knowledge space. In line with her background and her role some of the items cover topics in detail, other give just a superficial overview. Also in line with Ann's cognitive style her information space contains a large number of documents and an abundant list of topics. So she can really explore. Stepwise as she explores, her user model is updated and as a result of the continuous interaction between her personal knowledge agent and the three other agents her personal knowledge web extends and renews.

The power of intelligent tutoring systems lies in the fact that they are built on articulated models not only of what, but also of how to transfer knowledge. They are based on models of the domain, the student and on how to transfer knowledge. The weak side, however, is that the knowledge contained has to be prepared in advance and that the domain and user models contained cover only small samples of what is worthwhile or necessary to know. Knowledge management applications on the other hand and the techniques and tools used can

offer access to an unlimited amount of information. Internal business resources and access to the Internet infrastructure offer a theoretically unlimited access to the knowledge resources in a world without physical borders and can in principle be adjusted precisely to everyone's personal situation. The knowledge trees discussed offer a lifelong, student or personal model that can be carried as a personal knowledge passport through life. At each time and at each situation it is indicating what is known, at which level and from which role perspective, and the new knowledge targets to be achieved. At the same time knowledge trees offer possible links to persons with the same interest position. This paves the way for collaboration and collaborative support and is as such an important add-on that differentiates knowledge support systems from information systems.

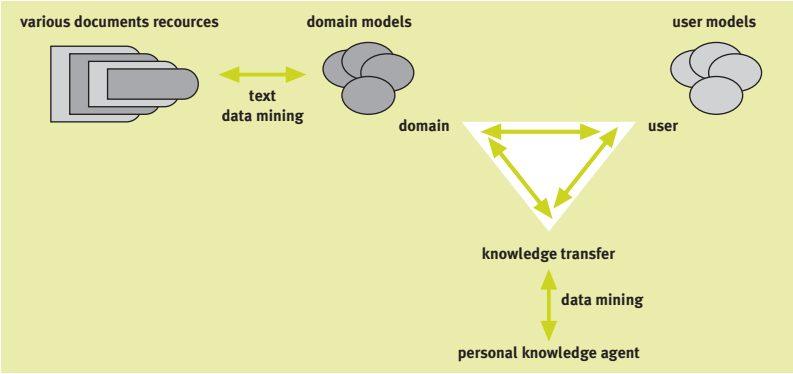


Figure 2
The knowledge triangle for personal knowledge management.

As a result, on the basis of the structure of intelligent tutoring systems, we propose a synthesis integrating the strength of these systems, knowledge management and the theory of knowledge trees. This implies a domain model and content (or content references) based on knowledge management techniques and tools, a user model based on the theory of knowledge trees and a knowledge transfer model based on intelligent tutoring systems.

Data mining and text mining techniques and applications will play an important role in supporting the domain, user and knowledge transfer agents to maintain their models. Their role in the implementation of the triangle for personal knowledge management is crucial. This will enable systems to move from a 'craftsmen phase' to industrial production both from a personalization perspective as well as from a content perspective.
Data mining techniques can be used to extend and maintain the user model. Text mining can contribute to the content of the knowledge base and create semantic maps for navigation (see Figure 3). The latter helps to structure even loosely structured content and to visualize relations the user has not previously been aware of. Thus, helping to transfer loosely structured and implicit (or tacit) knowledge into explicit knowledge which is the topic of the next paragraphs.

Creative processes, loosely structured and tacit knowledge

The starting point of the analysis is again the triangle model.

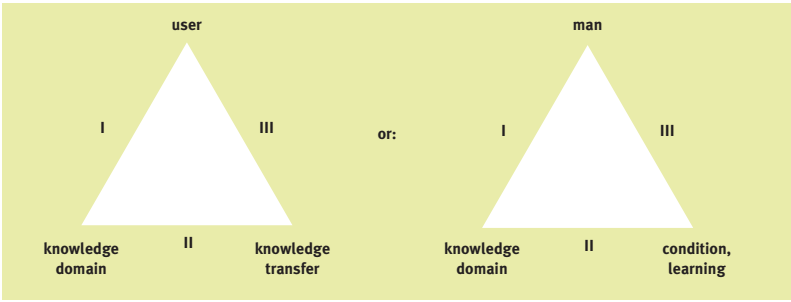


Figure 3
The 'knowledge triangle'.

When transforming information into knowledge the 'user' (which could be a person, group or even an organization) inherently uses the knowledge profiles acquired thus far in his/her knowledge domain as a starting point: the acquired knowledge profiles are combined and complemented with the (new) knowledge which the user is looking for, and which adds value in the user's personal knowledge management or in an organizational context. In Figure 3 this relation is represented by I; relation II deals with the methodologies, which can be used to translate knowledge concepts into knowledge, which can be learned by the user (line III).

Explicit and implicit knowledge; the concept of meaningful knowledge

At this stage it is important to distinguish between explicit and implicit (or tacit) knowledge.

Explicit knowledge is knowledge that has been codified or described in such a way that it can be transferred to a person in for instance a learning process.

Implicit knowledge is built out of a mix of personal experiences, skills and attitudes.

All knowledge used by a person is either tacit or rooted in tacit knowledge, which is illustrated by the 'iceberg' image of Figure 4.

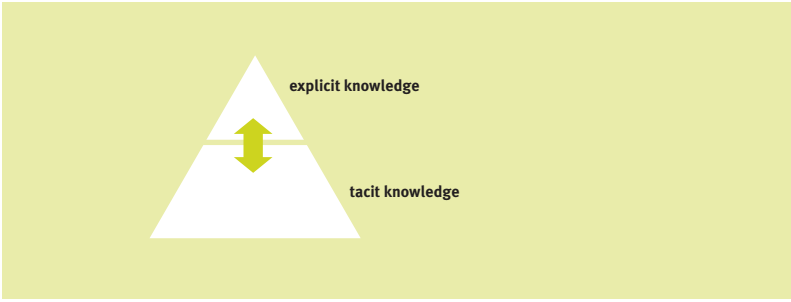


Figure 4
Explicit and tacit knowledge in relation to each other: explicit is rooted in tacit.

The tacit dimension of knowledge underlines the fact that knowledge is connected to a personal and social context: the act of knowing includes an appraisal and accordingly bridges the transition between subjectivity and objectivity.
The transition between explicit and tacit knowledge is subtle in the sense that it is a continuous one. This is illustrated in the matrix of Figure 5 [Tissen, 1998].

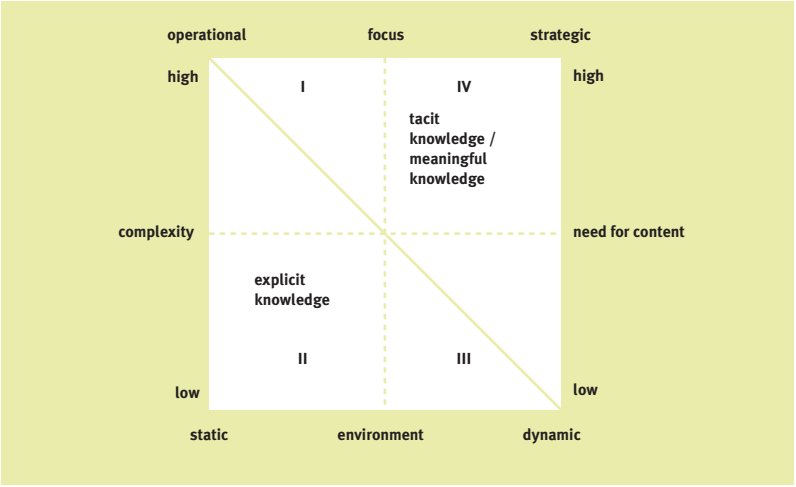


Figure 5
A matrix representation of the different types of knowledge.

An example of knowledge in quadrant I is knowledge used by legal persons; knowledge in quadrant II is fully operational information requiring less contextual information to understand, examples of which can be found in industrial production. Type III knowledge is used in a turbulent environment with a more strategic focus, e.g. in stock exchange. Quadrant IV represents implicit knowledge that is always meaningful to a person (or group), that is to say that has a specific meaning in a personal context, thereby adding value to the individual's life. Well-known examples are riding on horseback or knowing how to pour tea in a Japanese tea ceremony.
In this article we will not describe in detail how implicit knowledge is made explicit or vice versa. However, some basics have to be touched upon here and in the next paragraph.
First of all the lesson that everybody has learned in his or her life: physical exercises provide a deeper understanding than intellectual exercises, or — in other words — learning by doing is to be preferred.
What is behind this almost trivial notion is that the creation of knowledge that is meaningful to a person liberates (psychological) energy, it creates a positive emotional state. Think for instance of a 'Eureka experience' or the emotion coupled to grasping a complex relationship between knowledge concepts.
A second aspect of knowledge creation is the use of intuition. The intuitive capabilities of the human mind complement the rational way of thinking and are essen-

tial in solving problems or taking decisions in complex and strategic situations. When returning to Figure 5, it is apparent that tacit knowledge is extremely important in strategic situations with a high need for content in a dynamic environment. This is essentially the working ground for the (future) knowledge worker. In order to take decisions or reach conclusions the knowledge worker not only relies on facts, but also on intuition.
Working by using meaningful knowledge implies thinking in scenarios ('what if .....', or 'how could I .....'). Scenario-thinking is what people normally (should) do in all the divergent complex situations which they encounter in their personal life or in their professional activities.

One of the interesting things about scenario-thinking is that it is essentially based on 'both/and' logic. The implication is that you accept that a potential solution is not formulated in terms of X *or* Y, but in terms of X *and* Y. For instance, a new industrial product to be developed should not only have the required functional properties, but also comply with environmental standards and current legislation.
Scenario-thinking is about accepting creative, even in itself paradoxical approaches. The consequence of all this for the transformation of data into meaningful knowledge is discussed in the next paragraph.

Transforming data into meaningful knowledge

Transforming data into meaningful, implicit knowledge and also making this knowledge explicit, which is a prerequisite for effective knowledge transfer, will be the great challenge for persons, groups, and professional organizations, especially with a view to developing personal competencies and to life long learning.
First of all, it is necessary to have a short view at the mechanisms and techniques that govern the transformation between explicit and tacit knowledge (in both directions). In practice, three key factors support a successful conversion process [see also Ikujiro Nonaka, 1999]:

- expression and recognition;
- combination;
- integration.

Expression and recognition

Text data should be expressed as clearly identifiable concepts or figurative language (such as analogies or metaphors) in order to be suitable for conversion into tacit knowledge. A person is only able to 'grasp' the idea, to reach a real understanding, if he or she can visualize conceptual information by making the proper links to already existing tacit knowledge in his or her mind. Figurative language and the use of symbols are especially important, since a prominent

part of the development of (personally) meaningful knowledge takes place in an intuitive way in the subconscious domain of the mind.

Combination

Depending on the complexity of the data that are offered to a person, he or she will combine these data in the form of associated concepts (associations between the external data or with concepts already existing in the person's knowledge profile), which will build a concept network in the person's mind. Such concept networks generally have a unique individual character, since the association of concepts is a process that is determined to a large extent by personal experiences and fostered by their attached emotional value.
This directly implies that the emotional value that a person attaches to an association can spread in a continuum between low and high.
In the process of learning and of interaction between individuals in general, associative links may easily change to other concepts, together with the personal value attached.

Integration

As we have seen in the last paragraph real understanding is promoted in learning by doing situations. Often this requires interaction with another person (a colleague or teacher) or in a group. In this way knowledge is literally embodied in action and practice, and an integration process — as a consequence — takes place in the person's mind. The acquired knowledge then really becomes meaningful.

State of the art

We will now proceed with an overview of the state of the art, describing developments and tools dealing with the conversion of text data into meaningful knowledge, wherever possible referring to the key factors and mechanisms described above.

As we have seen, the first step to be taken in order to achieve a successful transformation of text data into knowledge is expression and recognition: text data should be expressed as recognizable concepts. Our mind associates a concept that is recognized in external text data with one or more concepts already present (as tacit knowledge) in the mind.
A semantic association in the form of a mental hyperlink is constructed. In this context a number of developments and IT-based tools are important. We will focus on:

– Algorithms for the self-organization of linking concept patterns in the Web to the patterns of an individual user, and self-organizing learning webs for knowledge structuring.

– Knowledge management systems enabling automatic categorizing, linking and delivering unstructured information from multiple input sources in such a way that users can efficiently locate and analyze the information they need.
– Tools for structuring knowledge, making visible links and associations between concepts.

Algorithms

Algorithms that will enable linking patterns of concepts in the Web to the patterns of an individual user are studied in cybernetics and artificial intelligence studies (see [Learning webs, 2001]). The basic idea behind several of the algorithms studied is that web links are similar to associations in the human brain, as supported by synapses connecting neurons. The strength of a link, like the connection strength of synapses, can change depending on the frequency of use of the link. This allows the network to 'learn' automatically from the way it is used. Such associative networks are more flexible than semantic networks (as used in artificial intelligence); they allow the expression of 'fuzzy' or 'intuitive' relations between concepts and have been regularly suggested as models of how the brain works. Algorithms for associative hypertext networks allowing 'self-organization' into simpler, more meaningful and more easily manageable networks have already been developed [Heylighen, 1999]. They almost allow us to think in terms of the 'Global Brain' metaphor, the development of the (currently static) World Wide Web into a self-organizing, thinking and dynamic Web. And even when a global brain cannot be reached, semantic web and search technologies will govern the future of the knowledge worker when he or she needs access to information.
We expect that such developments will be of growing importance for text mining and personal knowledge management.

Knowledge management systems

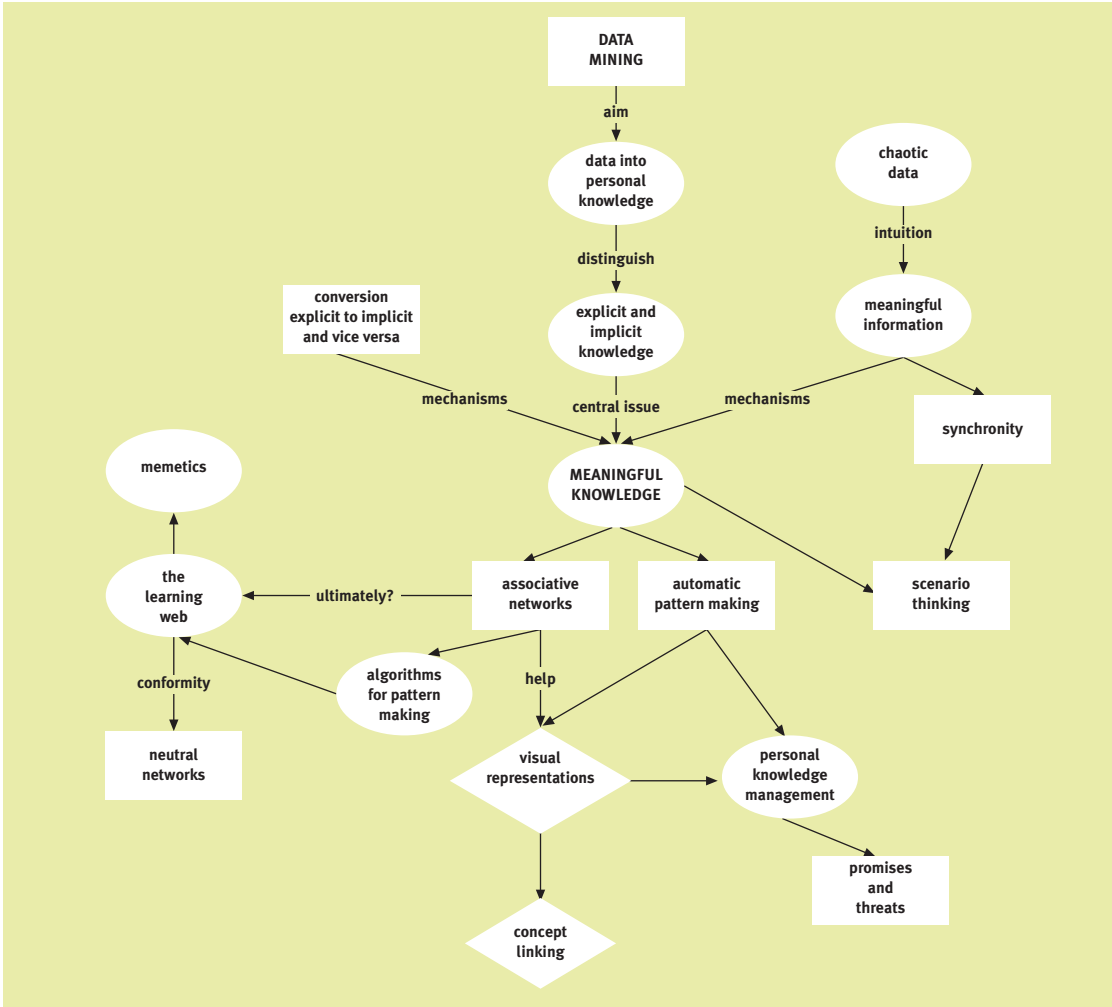
Already working examples of adaptive pattern-making algorithms are being used in knowledge management practice (see [Autonomy, 2001]). They categorize and link unstructured or loosely structured information from multiple input sources (the Web, intra-net sites, digital documents, e-mail), identifying the areas of expertise of individuals within an organization.
Future developments in this field include the widespread use of Adaptive Probabilistic Concept Models (APCM), as originally developed from research into neural networks. Here pattern-matching algorithms based on APCM are used to analyze documents and identify key concepts, which can be recognized by concept agents that automatically monitor the documents selected by users to view. Knowledge management systems, developed as in [Autonomy, 2001], use a person's reviewing behaviors to build a personalized profile of interests, which can then be used to achieve key knowledge management objectives.

Tools for structuring knowledge

As soon as (text) information has been drawn from input data, a person should keep track of his or her acquired knowledge in a personal knowledge management system. Most knowledge workers start by representing their ideas and knowledge in the form of a concept diagram or concept map. Nowadays, a growing number of professionals are using IT-based tools to make mind maps, fishbone diagrams (like in Figure 1), decision trees and concept maps or ontologies (specifications of concepts and tasks that define the knowledge and the system in which the knowledge base is described).

When making a visual representation of your own ideas you can recall the details better and you can recognize your gaps in understanding better. Since many people mainly think and learn visually, visual tools like, for instance, those used in Figure 1 and 6 are being increasingly adopted (see [Mindmanager, 2001]). Moreover, visual representation of associated concepts 'opens the gate-

Figure 6
Example of a representation of an associative concept network (the gross structure of the 2nd part of this publication in a two-dimensional image).



way' to our creative and intuitive abilities and thus facilitates the transformation of explicit knowledge into implicit, meaningful knowledge.
Web-enabled tools visualizing associative networks (as in [The brain, 2001]) will undoubtedly be the starting point of a number of developments directed at building personalized knowledge management systems based on associative networks of concepts (including all relevant hyperlinks).

Visions of the future

Managing 'chaotic' data: thriving on chaos

A lot of information comes to us in a 'chaotic' unorganized way: we receive information we did not ask for, we find web sites or references we did not specifically look for. Nevertheless, we often find very useful — even future determining — information in this 'chaotic' way. Moreover, the use of multi-reality scenarios in e.g. business situations includes the conscious search for meaningful information that comes to us seemingly by coincidence. We intuitively feel when new information is important, since only we can grasp the context. That is the reason why it is very unwise to rely on IT-based tools for filtering of information only: in our attempts to filter out the 'noise', we filter out the music as well. We have to use our human brain computer in a parallel, intuitive way.
On the other hand as stated in the beginning of this article, the world globalizes and the amount of accessible information grows at a dazzling speed. Therefore tied to our triangle model we foresee ICT and (text) data mining to have an important supportive role. This will enable us to identify topics of interest outside our personal neighborhoods.
The model we propose will enable knowledge workers to identify colleagues around the world who deal with similar or related topics. Their profiles will be similar or overlap and can (after consent) be used to identify peers. Once the contacts are made, interactions may lead to sharing and discussing new ideas or old ideas with new viewpoints.

Information only works when the two other building blocks of knowledge are present: intellect and interaction. The ancient Chinese already recognized this in their famous 'Book of Change' [Wilhelm,1950]. The basic idea behind their philosophy is that (emotionally loaded) thoughts and events (e.g. information one is provided with) are linked in a non-causal, but meaningful way. A micro-world event reflects the whole of nature and society and includes, within itself, the observer. Or, in other words, within each process of nature the whole is enfolded. The information that a person is looking for and which is meaningful to him or her significantly relates to patterns of change. The concept of synchronicity that summarizes this philosophy in one word was introduced by C.G. Jung (see e.g. [Peat, 1987]). Next decades will show a growing number of know-

ledge workers who recognize the importance of their creative and intuitive powers in retrieving meaningful knowledge from external data and who will not exclusively rely on information technology for the transformation of data into knowledge, and for taking important decisions based on them [Agor, 1986].

The management of implicit, meaningful knowledge

As we have seen, implicit, meaningful knowledge plays a central role in the transformation of data into knowledge. In the last decade a lot of attention has been given to knowledge management, especially in those organizations where knowledge workers have a prominent position. It is expected that managing implicit knowledge, both on a personal and on a group level, will gain primary attention in the next decades in relation to organizational learning and competence building. Models coming from the domains of change management and innovation diffusion will be applied to the management of implicit knowledge. These models address the psychological, social as well as cultural dimensions of acquiring knowledge [Nabeth, 2001].

Another area of potential importance is the study of memetics. A meme is defined as an information pattern, retained in an individual's memory, which is capable of being copied easily to another individual's memory. The most powerful medium for meme transition is the computer network [see Memetics, 2001].

Promises and threats

The use of personalized knowledge profiles, which describe 'gaps' in the knowledge domain of an individual, is evidently an important step forward in the life long learning perspective of the knowledge worker. But it also imposes a threat that this personalized information is being studied and used without explicit consent: autonomous user profiling of knowledge should be out of the question. Also, the past reviewing behavior of a person should not be taken as a basis for decisions to filter out information and data, which apparently do not fit into the personal knowledge domain. This also raises the possibility of multiple and 'marketed' identities of persons (with regard to information searching). A way should be found to determine the potentialities of both competencies and interests of a person, and based, on that basis, assist him or her in data mining. Moreover, technology should ensure people's (and organizations) rights to privacy and anonymity in this respect [Cingil, 2000; Pisa, 2001]. Private accessibility could be ensured by e.g. relevant combinations of biometrics and digital signature. But even then, ample room should be left for individual choices enabling self-tuition. Growth and development of an individual are to a large extent self-organizing, since the acquisition of meaningful knowledge takes place in a chaotic, self-organizing way.

The possibility of complete self-tuition will have an ethical side as well: making

children out of adults by stressing 'wish technology' and getting lost in fantasy worlds, developing hide and seek behaviors for fun, etc.
Another obvious threat is manipulating the news. The higher the degree of automation, the more sensitive the user will become to deliberate manipulation by hiding, inserting or promoting specific knowledge items.
Finally, we would like to stress the risk of exclusion. While the techniques proposed can, once implemented, be duplicated at relatively low costs, practice may be different. Personalized knowledge management should not lead to a situation where the already existing gap between rich and poor is enlarged, since only a small group of the privileged is assisted in finding or has access to the techniques and tools described in this publication.

References

- Agor, W. (1986). The Logic of Intuitive Decision-Making; a Research-Based Approach for Top Management. Quorum Books
- Albert, G. Mining Methods II. http://www.hindubusinessline.com/2000/07/26/stories/242639b6.htm
- Autonomy, 2001. Functional Review of Autonomy. In: http://www.doculabs.com or Algorithms and Methods described in http://www.inxight.com
- Bair, J., E. O'Connor. (1998). The State of the Product in Knowledge Management. Journal of Knowledge Management **2** (2)
- Brusilovsky, P. (1999). Adaptive and Intelligent Technologies for Web-Based Education. In: C. Rollinger, C. Peylo (eds.) Künstliche Intelligenz. Special Issue on Intelligent Systems and Teleteaching **4**:19-25. http://www.contrib.andrew.cmu.edu/%7Eplb/papers/KI-review.html
- Cingil, A., D. Azgin, A. Azgin. (2000). A Broader Approach to Personalization. Communications of the ACM - Personalization **43** (8)
- David, P.F. (1987). Synchronicity: the Bridge between Matter and Mind. Bantam Books
- Gruber, T. What is an Ontology. http://www-ksl-svc.stanford.edu:5915/doc/frame-editor/what-is-an-ontology.html
- Heylighen. F. (1999). Bootstrapping Knowledge Representations: from Entailment Meshes via Semantic Nets to Learning Webs. International Journal of Human-Computer Studies. http://www.mindmanager.com. http://www.thebrain.com
- Jasper, R., M. Uschold. (1999). A Framework for Understanding and Classifying Ontology Applications
- Lawson, J. (1999). Text Data Mining Introduction. http://allen.comm.virginia.edu/jtl5t/index.htm
- Learning Webs. In: Principia Cybernetica Web. http://pcp.lanl.gov/LEARNWEB.html

- Lehn, K. van. (1988). Student Modelling. In: M.C. Polson, J.J. Richardson. Foundations of Intelligent Tutoring Systems. pp55-78. Lawrence Erlbaum Associates Publishers
- Memetics. In: Principia Cybernetica Web. http://pcp.lanl.gov/MEMES.html
- Merlyn, P., L. Valikangas. (1998). From Information Technology to Knowledge Technology: Taking the User into Consideration. Journal of Knowledge Management **2** (2)
- Mizoguchi, R., K. Sinitsa, M. Ikeda. (1996). Task Ontology Design for Intelligent Educational/Training Systems. Position Paper for ITS'96 Workshop on Architectures and Methods for Designing Cost-Effective and Reusable ITS's. Montreal. http://advlearn.lrdc.pitt.edu/its-arch/papers/mizoguchi.html
- Murray, T. (1999). Authoring Intelligent Tutoring Systems: An Analysis of the State of the Art. International Journal of Artificial Intelligence in Education **10**:98-129. http://www.cs.umass.edu/%7Etmurray/papers/ATSummary/AuthTools.html)
- Nabeth, T., C. Roda. (2001). K-InCA: An Organisational Learning System which Exploits and Supports Social Processes. Publication INSEAD/CALT
- Nonaka, I., K. Noburo. (1999). The Concept of 'Ba': Building a Foundation for Knowledge Creation. The Knowledge Management Yearbook 1999-2000. Butterworth Heinemann
- Paiva, A. (1996). Communicating with Learner Modeling Agents. Position Paper for ITS'96 Workshop on Architectures and Methods for Designing Cost-Effective and Reusable ITSs. Montreal
- PISA. (2001). PISA - Privacy Incorporate Software Agent, Building a Privacy Guardian for the Electronic Age. http://www.tno.nl/instit/fel/pisa/
- Pouts-Lajus, S. (1993). Knowledge Trees, A Global System in Tribune. European Learning Technology Collection
- Scenarios for Ambient Intelligence. (2001). ISTAG Report. IPTS Seville, Spain. http://www.cordis.lu/ist/istag.htm
- Schmitz, M. Final Thesis. School of Engineering - University of Maastricht (in Preparation)
- Smith, B., P. Cotter. (2000). A Personalized Television Listings Service. Communications of the ACM - Personalization – **43** (8)
- Stoner, A. (2000). Knowledge Discovery and Data Mining. http://www.biermans.com/culminating/datamining.htm
- Syed, J. (1998). An Adaptive Framework for Knowledge Work. Journal of Knowledge Management **2** (2)

– Tissen, R. (1998). Value-Based Knowledge Management – Creating the 21st Century Company: Knowledge Intensive, People Rich. Addison Wesley/Longman
– Tiwana, A. (2000). The Knowledge Management Toolkit. Prentice Hall, USA
– Wenger, E. (1987). Artificial Intelligence and Tutoring Systems. Morgan Kaufman Publishers, California
– Wilhelm, R. (1950). The I Ching Book of Changes. Princeton University Press

5

5.8 Conclusions

Jeroen Meij

Data

Increasing data collection, digitization and application of intelligent software will contribute to the creation of large quantities of data about the individual. From technical developments in applications alone, we may expect a large increase in surveillance data, web mining data and survey data about individuals. Some of these applications will have to be subject to political debate to determine the actual level of proliferation desired.
The same activities will also be responsible for the availability of large quantities of data for individual use. People will create their own digital collections of literature, web pages, articles, music, pictures and video.

Data conservation

Anyone who has failed to find those important web pages which were available some months ago, realizes that the Internet is constantly changing. New pages are added, while other pages are changed or removed. Individuals or organizations depending on specific information need guaranteed access. This would mean either a public (national, global) Internet archive, or the creation of a private or local archive that contains important information. Several organizations and national libraries are involved in Internet archives. However for reasons of accessibility and control, local archives (intranets) are expected to be the common solution. The national libraries will be historical archives, documenting the evolution of the Internet.

For any electronic archive the ageing of carriers and data or operation system formats is an important issue. Possible strategies include regular migration to other carriers, emulation and encapsulation. However, this issue is still largely unsolved. For documents standardization is an important topic, but there is no final solution yet, since the standards are continuously evolving.

Multimedia mining

The international research community has become aware of the need for multimedia mining. In the next decades, we can expect real multimedia mining applications to enter the commercial realm. This will be made possible through two developments:

- Dedicated algorithms. In present multimedia mining practice, generic data mining algorithms are applied to multimedia data. In the near future we will see a drift towards new multimedia mining dedicated algorithms, algorithms specifically designed for the high-dimensional, rich and complex multimedia data. Developing these algorithms will be essential to movement across the present barriers of multimedia mining.
- Close collaboration between data-, sound, video, image processing experts, usually working on relatively small data sets, and data miners, focusing on large data sets. This collaboration should not only be vertical, but also cross disciplinary to enable true multimedia mining applications.

Image

Content-based image retrieval and classification has reached a mature state and various commercial products have been brought onto the market.
From a scientific perspective the following trends can be distinguished. First, large scale image databases are being created. Obviously, large scale datasets provide different image mining problems than rather small, narrow-domain datasets. Second, research is directed towards the integration of different information modalities such as text, pictorial, and motion. Third, relevance feedback will be and still is an important issue. Finally, invariance is necessary to get general-purpose image retrieval.
Indexing, searching and assessing the content of large scale image databases will be done by software tools, not by humans.

Video

We can anticipate that further development of video-content analysis algorithms will strongly accelerate in the years to come. Video mining will be an important tool to deal with the multimedia data overload, with video of particular interest considering the quantities of data and its popularity. Nobody dares to claim that the semantic gap will be bridged in 20 years. Product suites for content-based image and video indexing and searching will be developed.

These tools will serve the needs of future content owners in the field of entertainment, news, education, video communication and distribution.

Music

Due to the great variability of musical audio, its non-verbal basis, and its interconnected levels of description, musical audio mining is a rather difficult field and still in its initial stage of development. Musical audio mining is rooted in musicology, where it draws on concept taxonomies that allow users to specify a musical piece in terms of more or less unique descriptors. These descriptors have their roots in acoustical properties of the musical audio, hence signal processing and statistical modeling are core disciplines as well, because they relate the audio to the conceptual taxonomy. As in text-based data mining, similarity measurement plays an important role in finding the appropriate connections between representational structures in the query and representational structures in the database.
In the next stages of development, musical audio mining products will be employed by professional content distributors, entertainment and leisure industry and, finally, by the consumer.

In the very near future, we will see multimedia data mining tools as conventional applications in cars, in homes, and even with wearables (i.e. computer powered cameras, built-in garments). Cameras mounted on computer displays could identify user emotions and interpret needs. Identifying and recognizing objects in real time will become common practice. Cameras mounted on mobile carriers, such as cars or even humans, will have enough computing power to help users recognize and interpret the environment in which they proceed. Such devices could help car drivers in tracking potentially dangerous situations or warning the driver of fatigue or distraction.

Text mining

Speech

Although not discussed as a separate subject, speech analysis is a very important step for achieving context-aware systems. The speech step in these systems will be likely to be a speech recognitions system, converting the speech to text. After this conversion, the analysis will be a text mining step.

Text

Most current text mining applications are limited to index and retrieve functionality. Only few tools currently go beyond the document retrieval stage. The next challenge in text mining technology will be the interpretation of a text, conveying the message the authors wanted to expose.

The combination of current and emerging machine leaning technologies will impact on the way people work. Combining text mining and data mining technology with general machine learning technology will yield a next generation of intelligent adaptive knowledge management systems. These knowledge management systems will be able to increase their knowledge of the domain with the growing number of documents contained in the system. Moreover the adaptive knowledge management system will be able to adjust its knowledge, when documents from new domains are added to the system. Especially for multidisciplinary and fast changing markets, such as the professional services organization industry, the next generation knowledge management systems will be of interest.

Web mining

The World Wide Web, viewed as a huge and heterogeneous repository of information, motivates the development of new techniques for retrieving information, techniques that are much more sophisticated than the ones typically used for classical databases. There is cross-fertilization between information retrieval and extraction on the one hand, and data mining on the other hand. Both may be useful as a component of the other. The state of the art in both domains is steadily advancing, as can be seen by several impressive applications that already exist on the Web. On the assumption that the current trend continues, it is reasonable to expect that in the next decade the Web will evolve into a knowledge base, the completeness and intelligence of which will largely surpass that of any encyclopedia, newspaper or classical library, and, for many domains, even that of human experts.

Web–content mining

The techniques described in this section can be used to automatically build a giant knowledge network by analyzing a very large number of documents. In turn, this knowledge network enables a search engine to cluster the search results into meaningful categories, which assists the user in finding the relevant information. Furthermore, this knowledge network and the techniques to generate it, can of course be used for a myriad of other applications like document classification, topic determination, document comparison etc.

Web-adaptive technology

After considering the diversity of methods and techniques for adaptation we come back to the crucial role that user modeling plays in this process. In order to be able to have an adequate process of adaptation, it is essential that the right data is available to base the adaptation on. The analysis of previous usage data is one of the ways to achieve this. These usage data, as stated earlier, appear in a format that contains a lot of detail, but that does not offer a high-level view of the user behavior.

While some advantage can be obtained by the collaboration of the users, either individually or collectively, that approach does not suit the needs of all applications. Some applications cannot place this burden on their users. Therefore, in these applications, usage data is the source of information that should be exploited.
It is exactly for this reason that data mining can play a role in adaptive hypermedia. The future developments of adaptation can therefore benefit from further developments in data mining.

Knowledge integration and learning

The use of personalized knowledge profiles, which describe 'gaps' in the knowledge domain of an individual, will be an important step forward in the life long learning perspective of the knowledge worker. For best results, a way should be found to determine the potentialities of both competencies and interests of a person, and based on that to assist him or her in data mining. But even then, ample room should be left for individual choices enabling self-tuition. Growth and development of an individual are to a large extent self-organizing, since the acquisition of meaningful knowledge takes place in a chaotic, self-organizing way.
With regard to personalized knowledge profiles, technology should ensure people's (and organizations) rights to privacy and anonymity. Also, care should be taken to prevent unequal access to these tools across society, which may worsen the digital divide.
An interesting observation is that when organizations become learning organizations, increasing the learning rate (as is to be expected), the focus of knowledge management could shift from externalizing implicit knowledge towards managing and disseminating the knowledge worker's learning processes and the involved explicit knowledge.

Future knowledge workers

Now, can we envision the changes that will occur when mature video, audio, text and multimedia mining tools are commercially available? Where agents know our behavioral patterns and will react on what we experience and anticipate on what we want?

From the developments sketched in this part, we can formulate a vision of future knowledge workers. When we combine such a general vision with three different profiles of knowledge workers, we might see different levels of intensity of use:

– The first group of people are permanently connected to the internet, but also have their own data repository. They are assisted in performing their tasks by advanced search, analysis en presentation (summarization, graphics) soft-

ware. Software agents continue gathering, while they are doing other things. Input is mainly text- (typed or voice) or graphic interface based, output on a screen or head mounted display. Even now, for specific applications, augmented reality is already possible, for instance aircraft mechanics can project data about the parts they are working on over the image of a specific part on their retina. In this way their working environment is enriched.

– Interacting in a more intense way, another group of people will be immersed in their search and productive environment at times, where interaction with advanced tactile and movement sensors and actuators enables them to explore and act. These immersion techniques will also make virtual presence en cooperation possible. Surgeons in different physical locations will discuss 3D scans of patients, while navigating through the 3D images, manipulating the representations of various tissues to determine optimal procedures.
– The highest interaction intensity will be reached by those who will be connected through clothing, accessories and implants[1] during a large part of their day, working in a highly augmented reality. Context aware agents supply them with additional information on their physical or search environment continually. Besides having the advantage of being well informed at all times, they probably will have short periods of absentmindedness, when they are communicating with the system or with each other.

These examples may seem far away from present everyday reality, however in laboratory settings in dedicated tasks many things have already been demonstrated.

1 See for example http://homepage1.nifty.com/konomi/shinichi/ubicomp.html

6

Methodology and Technology

6.1 Methodology

Maarten van Someren[1], *Arno Siebes*[2], *Floor Verdenius*[3], *Jeroen Meij*

6.1.1 Introduction

This part of the book will provide a technical perspective on knowledge discovery and data mining. We will start by discussing some of the terminology in Section 6.1.2, followed by a historical background of the technology (Section 6.1.3).

We can make a division of the knowledge discovery process in the following steps [Adriaans, 1996; Witten, 2000; Fayad 1996]:

- Problem analysis.
- Data acquisition.
- Data processing.
- Data analysis.
- Reporting.

Though various authors use slightly different terms, the general idea of common subdivisions corresponds to that given here. The steps are discussed in Section 6.1.4.

1 Dr M. van Someren, maarten@swi.psy.uva.nl, The Universiteit van Amsterdam, Faculty of Social and Behavioural Sciences, Department of Psychology, Amsterdam, The Netherlands

2 Prof Dr A. Siebes, arno@cs.uu.nl, Institute of Information and Computing Sciences, Department of Mathematics and Computer Science, Utrecht University, The Netherlands

3 Drs F. Verdenius, F.Verdenius@ato.wag-ur.nl, Department of Production & Control Systems, ATO, Wageningen, The Netherlands

Of course, this process is embedded in one or more tasks, which in turn are embedded in goals and targets. This process of embedding is discussed in Section 6.1.5, including some thoughts on the selection of techniques for the analysis phase. Also the interpretation and evaluation of the results is considered. The technical integration in the IT environment closes this chapter with Section 6.1.6.
The techniques for analysis will be covered in Chapter 6.2.

6.1.2 Some definitions

Data mining

In the context of this project, we propose to amend the common definition [Fayad, 1996], not restricting its use, and introduce the term aggregation [Hand, 2002, in Chapter 1.1 of this book]. This leads to the following definition of data mining:

Data mining is the process of extracting previously unknown information from aggregations of data. In the right context, this leads to knowledge.

The primary interest in data mining is aimed at creating representations that reflect relations within the data, rather then just filtering part of the data when doing a query.
This representation is in some cases the final product of the analysis step. However, it is often used for a further goal such as the prediction of future events, recognition, construction or optimization of new descriptions. The representation can be in the form of a description in language, a numerical or other formal model, or a graphic representation. It can be implemented in an executable computer program that can be used for automatic prediction, recognition, construction, optimization, etc.
Data mining is related to statistical data analysis and numerical modeling.

Knowledge discovery in databases (KDD)

KDD views data mining as a subprocess. It includes more of the methodology around the technical data mining (e.g. problem analysis, data acquisition, presentation of results, etc.). On the other hand it does not include the use of data mining for the construction of systems. Machine learning includes data mining, including its use for the automatic construction of systems. The main focus in machine learning is currently on methods, algorithms and their properties and less on the embedding of these methods in systems.

Data, information, knowledge

To prevent confusion between the terms used above, I propose to use the following definitions [Schreiber, 1998]:

- Data are the 'uninterpreted' signals or symbols, that reach our senses, for example a red light (lamp) near a crossing.
- Information is data with a meaning attached to it, for example a red traffic light indicates 'stop'.
- Knowledge is the whole body of data and information that people convert to action, in order to carry out tasks and to create new information. The knowledge level adds a sense of purpose and a generative capacity. The driver

approaching a red traffic light is aware of the purpose of traffic regulation and of the possible consequences, if he does not stop.

Table 1 summarizes the differences between the terms.

Table 1
Data, information, knowledge.

	Characteristics	Example
Data	Uninterpreted, raw	...—-...
Information	meaning attached to data	S.O.S.
Knowledge	attach purpose and competence to information, potential to generate action	emergency alert-> start rescue operation

Knowledge is very dependent on the observer and the context: one man's knowledge could be just data to someone else.
For our purposes, the most important thing is that the user wants information, not data. This Information should be presented in such a way that the user can convert it into knowledge.

Information retrieval (IR)

IR is concerned with the problem of retrieving a document from a large set of documents. The retrieved document(s) should satisfy a query, a description provided by the user. Since no previously unknown information is discovered from the aggregation of data (all documents must be present in the data set to be retrieved), this is not considered to be data mining.
However, data mining can be used in several ways to support the performance of this task. For example, a model can be constructed from a collection of documents or can be used to construct a model of a user or a user population. A model of a document collection can be used to speed up retrieval or to help the user define a query. A model of a user (population) can be used to redefine the query. If the information that must be retrieved is not a complete document, but part of a document (or part of a set of documents) and the target information is described as a class rather than a query, then the task is called information extraction. For example, finding a document about data mining, computers and biology is an information retrieval problem. Defining a rule that extracts the biological variables from documents describing biological data mining studies is information extraction.

6.1.3 A brief history of data mining

The area of data mining has its roots in three areas: databases, machine learning and business administration. In the area of databases two issues lead researchers to consider the possibility of data mining. One question was the issue of compressing large databases. Regularities in a database allow the database to be stored in a more compact form without losing any information. This motivated the development of methods that search for such regularities. If these methods are allowed to find regularities that are approximate they amount to data mining methods. A second issue was a generalization of retrieval to more general database queries. Answering general queries about the relation between variables requires methods that search for general patterns. These questions led to the development of methods that search for 'association rules' and their implementation for large databases. Interest in these issues was reinforced by the use of database technology for 'business information systems', 'decision support systems', 'management support systems', etc. These systems store information about business processes. Initially, they were used to retrieve specific information such as the availability and cost of goods, but increasingly their potential for providing more general information was recognized, which led to the need for more powerful methods.

Machine learning was originally a branch of artificial intelligence with the goal of finding methods that reproduce human learning. The main focus of the field became the study of methods that construct general models of observational data. For a relatively long time machine learning has been a scientific discipline. Its links with industry were considerably weaker than those of database research. Also, machine learning research was biased towards non-numerical data and models. Numerical models were considered primarily an issue for statistics and applied mathematics. As a consequence, the potential of machine learning methods remained largely unused, until the notion of data mining appeared.

Classic statistical theory was mostly concerned with the problem of hypothesis testing and not with the problem of finding models for data. This was considered exploratory analysis that is more difficult to understand from a theoretical viewpoint. Statistical methods for testing a hypothesis are valid for a one-step testing process. If more hypotheses are evaluated on a single dataset, as is common in data mining, classic statistical theory cannot be directly applied. On the other hand, the non-classical Bayesian approach to statistics fits very well with data mining. In this model, prior knowledge (of distributions) is updated from new observations. The problems that were predicted by classical statisticians are now generally recognized in data mining and the collaboration between statistics and data mining is now increasing.

6.1.4 Process steps

We will focus on the last four steps of the data mining process given in Section 6.1.1, while problem analysis will be discussed in Section 6.1.5:

- Data acquisition.
- Data processing.
- Data analysis.
- Reporting.

Data acquisition

The data that will be used to identify interesting structures has to be acquired first. The character of this stage depends on the application context. One extreme is a situation in which nothing has happened yet and the entire process is planned in advance.

After determining the questions to be answered, the data has to acquired or selected. On one end of the scale, methods of data acquisition resulting in the desired content, quality and quantity of data can be planned and implemented. The other extreme case is one in which the data to be used is restricted to an existing data base.

Most projects fall between these extremes: a source of data is available, but which data are useful and which will be used is still open. Or there is a possibility of collecting certain data, but the actual acquisition still needs to be planned. Most data mining methods concern the analysis of data and little is said about acquisition. The main problem for planning data acquisition is to balance the potential relevance of data against the costs of acquiring them. Both the nature and the amount of data must be planned and the costs must be compared with the expected costs and benefits of alternatives to data mining [Verdenius, 1999]. The price of not recording relevant data is the discovery of fewer patterns in the data. The price of recording too many variables is in the acquisition costs: more variables must be measured, recorded and stored. Furthermore, additional variables need additional observations to avoid spurious patterns. As a rule of thumb we can indicate that n variables need 10n to 100n observations.

More variables also increase the cost of analysis: it takes more time and it may need additional work to make the problem manageable by a computer. For some techniques the computational cost (CPU time) increases exponentially.

When a large database is available, a planning stage may be required to decide which of the available data are to be used. Planning may also identify existing knowledge about the domain, which can be used in combination with inductive methods.

Data processing

Cleaning

It is almost inevitable in realistic situations that data are incomplete (for some objects or events some variables have not been recorded) and contain errors. Some methods are available for data cleaning:

- *testing integrity constraints*: defining the possible values of single variables or possible combinations of pairs or even triples of variables and checking, if all values or value pairs satisfy these constraints.
- *visual inspection:* displaying the distributions of variables (or the joint distributions) and asking a domain expert to check, if these make sense. Data that the expert rejects are removed or changed manually.
- *unification*: if the data come from different sources, they need to be unified on the basis of a common conceptual model or 'ontology'. For example, two hospitals may use different terminology or different concepts to describe patients. Merging their data requires the design of a uniform language for describing patients and also the translation from 'raw' data to the unified form.

Enrichment

Data can be enriched with additional observations or other databases, adding new variables to the original data set.

Transformation

There are several reasons for transforming data prior to the actual analysis. Assuming that the cleaning phase resulted in data that will be explored by the actual data mining methods, it is sometimes necessary to transform the data. We can distinguish source related and technique related transformation. Source related transformation often involves 'discretization' of continuous data, for example from measurements.

Technique related transformation is often required because data mining tools take input in a certain form and construct knowledge representations of a certain kind. Data must be brought in a form that satisfies the input format of the analysis tool and that allows a tool to find a representation of a certain kind. Most data consist of a set of descriptions of objects or events. The description can include both numerical and non-numerical variables. It is important that the analysis tool 'knows' this, because patterns that involve numerical variables are different from patterns for non-numerical variables. For example, if the age, weight and disease of a patient are coded as numbers, then the data mining method may search for a numerical relation between age and weight, but a numerical relation between age and disease should not be expressed as a numerical function. Even if diseases are coded as category numbers, this only

means that they are different and it does not imply any ordering or underlying dimension.

Other forms of transformation are directed at reducing the length of scales (for example by introducing intervals for numerical values or grouping values of non-numerical variables), reducing the number of variables (by testing if they show any relation with other variables or by transforming the entire space with techniques such as principal component analysis, see Section 6.2.4).

Analysis

In the analysis phase, it is important to recall the goal of the analysis to determine, whether we want to learn something about our data base only (derive specific knowledge), or we want to generalize the findings to outside our database (derive general knowledge). The first case is applicable for example, when we want to describe a certain image collection or a closed collection of documents. The second case is very common, for example when we want to create a set of rules that can be used for prediction or classification.
To be able to generalize, we have two things to take into account for our knowledge representation:

- Prevent overfitting.
- Check statistical validity.

Over fitting describes a situation where the knowledge representation (model, set of rules) perfectly describes the data set, including accidental patterns that only occur in this data set. Any 'overfitted' rule applied to data outside the data set is likely to produce erroneous results.
A common way to remedy this is the use of a training and a test set of data. The training set is used to derive the rules, the test set to validate the general applicability of the rules. See also Section 6.2.19, Boosting.
The statistical validity of the rules can be tested using the available hypothesis tests (see Section 2.2.4).

A commonly used division of analysis tasks is that of supervised/unsupervised learning.
Supervised learning means that a model is constructed that is to be used in combination with specific input to infer specific output. For example, learning a model that can predict the disease of a patient from symptoms is called 'supervised', when the data base already contains verified disease diagnoses. The searching of a database of patients for any kind of pattern is called 'unsupervised'.

Data mining analysis techniques

The range of analysis techniques is very wide and increases rapidly. The most important techniques will be discussed in Chapter 6.2. Usually, a technique as described in Chapter 6.2 can use different algorithms, the actual mathematical calculation methods, implemented in a computer code. Discussion of the algorithms is beyond the scope of this publication. However, sometimes we list the most important ones for a particular technique. The selection of techniques is discussed below and in Section 6.4.4, Meta-learning.

Technique selection

Floor Verdenius[4]

Often, data in a database has been designed to fit an operational goal. Then, when a proper data set is obtained that suits the problem characteristics, the technique or techniques have to be selected that can solve the problem with (an expected) satisfying performance. Unfortunately, technique selection is not straightforward. There is a poor understanding of data set characteristics and technique performance. Experimental studies [Michie, 1994] and also [Lim, 2000] deliver little grip on selecting the best technique. Current strategies include:

Using familiar and available techniques

Many data mining researchers have a small number of favorite techniques; based on their experience in tuning these techniques on specific data sets, their performance in using them is reasonable. When carrying out a project, this strategy invests effort in pursuing the ultimate result by squeezing the maximum out of a technique.

Experimental assessment

When many techniques are available in an experimental environment, all techniques can be run on the same problem. By comparing performances between techniques [Michie, 1994], the best performing technique can be selected. In this strategy, effort is invested in identifying the optimal technique.

Assessment on the basis of data characteristics

By calculating data descriptors of all kinds, and correlating these to performance statistics of many different techniques, determinant data characteristics for specific techniques may be found [Michie, 1994; Brazdil, 1994; Engels, 1998]. The outcome of such an approach supports a priori selection of techniques. In this approach, analyzing the data set to use consumes most of the effort.
In practice, all these approaches have advantages and disadvantages. A major drawback of the first two approaches is that there is no understanding of why

4 Drs F. Verdenius, F.Verdenius@ato.wag-ur.nl, Department of Production & Control Systems, ATO, Wageningen, The Netherlands

specific techniques perform well on specific data sets. The third approach does better in that respect. In general, there exists little guidance in understanding of the performance of specific techniques. Expanding the latter approach, [Verdenius, 1999] has analyzed the relation between specific data characteristics and the performance of specific techniques. The set up for this approach is that each technique has a guard, which tests whether a data set reflects statistical properties that underlie the technique at hand. Decision trees, for example, operate best on data that exhibit class boundaries which are orthogonal to attribute axis. The guard for univariate decision trees therefore assesses the existence of orthogonal class boundaries. If sufficient evidence for such boundaries is encountered, the guard allows the data to be analyzed by univariate decision trees.

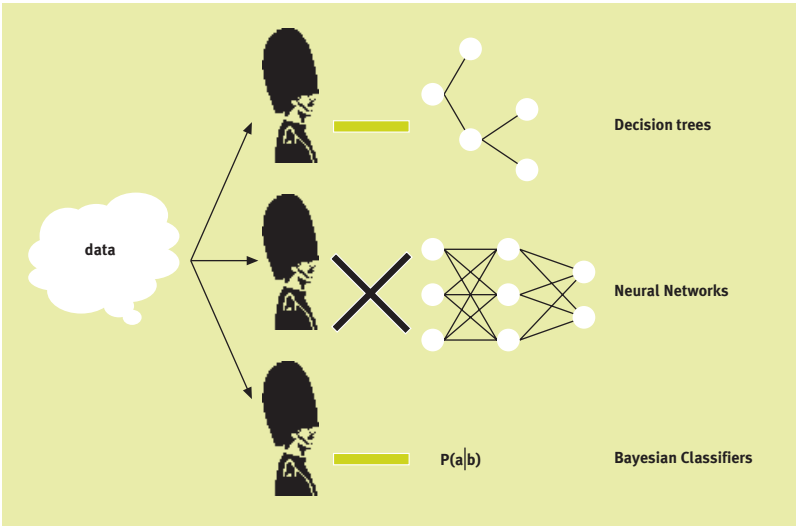


Figure 1
Rational technique selection by guards. A guard is a small software procedure that assesses the suitability of one inductive technique for analyzing a data set. The guard results in a quantitative assessment of suitability. For decision trees, such a guard has been described in [Verdenius, 1999].

Further tools for automatic technique selection are discussed in Section 6.4.4, Meta-learning.

Technique requirements

When a technique has been selected for a data set, the technique has to be tuned for that data set. This may include specific data transformation, additionally to transformations in the analysis level[5]. Techniques may impose specific demands on data representation. Specific neural networks, for instance, perform better, when numeric values are encoded in a special binary format (cf. Gray code), and certain decision tree and decision rules improve learning accuracy, when data is discretized. Now, the problem (data) is ready for training. Most techniques take a number of parameters. The values of these parameters may critically influence the performance of the technique. Finding the optimal values for these parameters is done during the tuning phase. For neural net-

5 It is important to separate problem related data transformation from technique related transformations. At the analysis level, data is delivered in a format that optimally represents the problem. Specific techniques may impose additional requirements on data representation, but these requirements should not influence the problem at hand.

works, such parameters may include design (number of layers, numbers of node per layers) as well as learning behavior (learn rate, momentum). For decision trees the pruning criterion (See Section 6.2.7) can be seen as such a parameter. Some techniques include automatic tuning. This can be some optimization routine like hill climbing (See section 6.2.6). For many techniques, however, tuning is a manual activity. The analyst performs a number of experiments to find the optimal setting of parameters. Extensive literature exists on how to compare performance of different models. One often used approach is cross validation [Michie,1994].

Reporting

After the analysis, some kind of report has to enable the user to access the results, the actual delivery of the knowledge representation. The form that this representation takes depends on the aim of the project. One extreme is that only the knowledge formulation is of interest; the actual decision tree, decision rules, or regression parameters or neural network weight. On the other extreme, not so much the actual model, but the implemented procedure to extract that knowledge from data is the desired result. It is clear that the latter imposes other (and more strict) constraints on analysis of the application goals and on the data analysis than the first aim.
Often, sets of rules are generated, each with a given support and confidence (see Section 6.2.1). Visualization plays an important role in the reporting phase (see Chapter 6.3).

Using the results of data mining

The result of data mining is a (partial) model of data. This can be used for very different purposes. We can distinguish the following main types.

Human inspection and analysis

The model is inspected by a person who uses it to extend her own model of a process, event or state. For this purpose a model must be in a 'communicable' form. Usually the models constructed by data mining tools stay quite close to the data and do not include interpretations. For example, a medical database may contain observations about patients and records of their medication, but no information about the processes and structures that explain the course of the disease. As a consequence this intermediate deep knowledge will not appear in the model and the user of the model must integrate it with her own deep and experiential knowledge.

Simulation and prediction

The model is used to (automatically) simulate the effect of changes. For example, a model of a production process that is constructed by data mining can be

used to predict the effect of (unobserved) changes in the input parameters, or a model of a disease can be used to predict a disease from symptoms or to predict the effect of medication on the development of the disease.

Control

Models can also be used to optimize performance. For example, the model can relate conditions of a process to optimal interventions and be used to control interventions in new situations.

References

- Adriaans, P., D. Zantinge. (1996) Data Mining. Addison-Wesley/Longman
- Fayyad, U.M., G. Piatetsky-Shapiro, P. Smyth. (1996). From Data Mining to Knowledge Discovery: An Overview. In: U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy. (eds.). Advances in Knowledge Discovery and Data Mining. AAAI-Press. pp1-37
- Witten, I., E. Frank. (2000). Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann

6.1.5 Process embedding

Floor Verdenius[6]

Introduction

Data mining as a discipline has gained popularity after a diverse suite of inductive techniques emerged from several disciplines. A large number of inductive techniques is assessed by [Michie, 1994] on various data sets. In the respective research fields of statistics, machine learning and neural networks, the main focus has always been technological. A technical expert trying to solve a real world problem tends to squeeze the maximum out of the available techniques in order to solve the problem. A problem that proves difficult to solve with existing techniques, often triggers the development of new techniques that solve the new problem (type) better, rather than looking over the disciplinary boundaries for other tools that can handle the problem.

With the availability of a broad suite of inductive techniques, data miners began to concentrate on structuring the process of applying these tools. As elaborated in the previous section, a stepwise approach to data mining analysis can be given: problem analysis, data acquisition, data processing, data analysis, reporting.

[Brachman, Anand, 1996] elaborate that view by stressing the importance of the function of knowledge discovery from data bases (KDD). The KDD process is one step in a comprehensive process to support a business function (e.g. marketing, process planning). KDD delivers not (merely) knowledge, but also the specification of an application that is to exploit the discovered knowledge.

Developing this specification requires both data mining expertise and profound knowledge of the application domain. Realizing the actual business application is the job of the more traditional system developers and knowledge engineers. In this view, we can expand the steps with some preceding steps and an implementation/application phase.

Two questions precede the data mining phases as shown in Table 1:

– *Is data mining the best approach for analyzing the problem at hand*? Before designing a data mining solution, it has to be assessed to what extent other approaches can play a role in solving the problem at hand. Available expertise, domain theories and models may be used for solving (sub)problems. Analyzing the overall problem, and assessing alternative (sub) solutions in the light of exploiting available data, when appropriate. The result of this analysis is a problem decomposition to the level of solvable subproblems.
– *How can we identify and acquire relevant data to solve the data mining problem*? In both aforementioned views, the available data in a data warehouse is the starting point of the analysis. In many domains, e.g. in agricultural environments, the actually recorded data is only a subset of the available

6 Drs F. Verdenius, F.Verdenius@ato.wag-ur.nl, Department of Production & Control Systems, ATO, Wageningen, The Netherlands

data. Moreover, depth analysis of the domain often reveals complementary data that, if made available, could make the application benefit in terms of the predictive and structural accuracy.

It is obvious that these two issues are not completely independent. The outcome for one issue constrains the possible outcomes of the other. Analyzing these points includes making choices, and may include small scale experiments, e.g. in order to assess added value of specific data attributes.
The data mining steps can be seen as a layered process (see Figure 1). Each layer has its own focus. The next sections will subsequently attend:

- the application layer, where the functional decomposition and design steps are located;
- the analysis layer, where detailed analysis of available data leads to technique selection and parameter design;
- the technique layer, where technique parameters are established, and the induced model actually is generated.

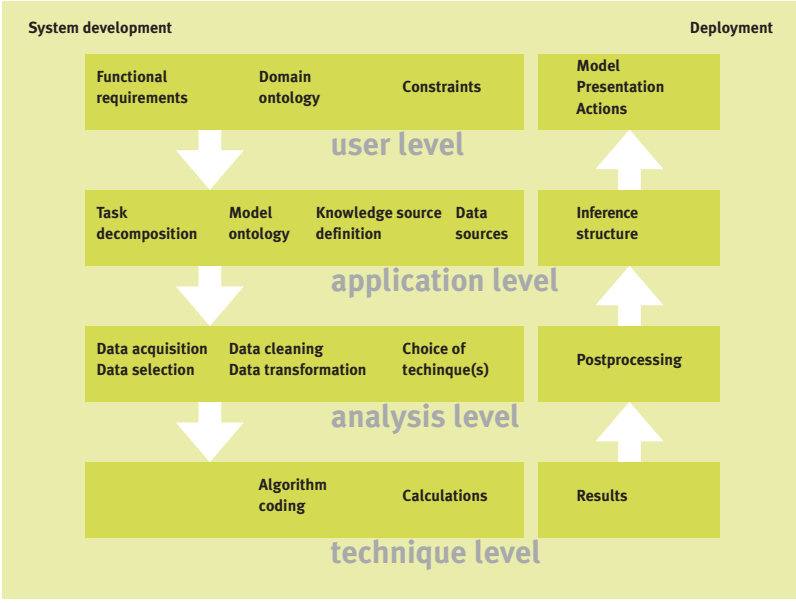


Figure 1
A layered model of a data mining process.(freely rendered from [Verdenius, 1997]).

User and application levels

For a data mining process, the requirements of the application in the form of a functional requirement (cf. knowledge goals [Hunter, 1993]), a domain ontology and non-functional requirements serve as starting point. At the application level these are transformed into a decomposition of the global functionality into (a number of) subtasks, and the required data to be used is defined. Apart from knowledge of the problem domain (for assessment of potential data sources and global requirements), this level requires adequate knowledge of the com-

petencies of different solutions (including knowledge technology and data mining algorithms), to be represented in a primitive task knowledge base.

Analysis level

At the analysis level, for each of the relevant subproblems data is acquired from the real world. This can be from an operational data warehouse, but the data may also be obtained in dedicated data acquisition projects. Data is assessed on data quality:

- Is the data complete and interpretable (in other words: do records have substantial attribute values for every attribute; if not: what are the characteristics of missing data).
- Is the data sufficient and representative for the domain (including a study of what representative means in the domain at hand).
- Data quality problems are repaired during data clean(s)ing. Records with quality problems (e.g. too much missing attributes, contradictory content, recording or measurement errors) are either deleted from the data set or repaired.
- Furthermore, data can be transformed to suit the problem at hand. Data transformation includes representation change, attribute extraction, selection and construction [Liu, 1998]. If a fruit firmness measurement delivers a value from a continuous range, this attribute may be discretized. Such discretization may be defended from the application domain (during the ripening of fruit, a climacteric phase causes a sudden and manifest decrease in firmness, but soft fruit measurements are inherently unreliable with the available instruments). Another type of support for such a discretization may come from a data analysis observation showing a class boundary in the domain of an attribute.

Application modes

In general, several modes of application can be distinguished. First, the aim of a data mining project can be the formulation of knowledge from data. Business tasks like clustering and classification (see Part 3) fall into this mode. In a project on fruit treatment [Verdenius, 1996], the need emerged to preselect fruit in order to make a measurement more informative. Product experts, the actors going to perform that selection, had no knowledge of how to recognize the right fruits. A large experiment was set up to measure all (potentially) relevant features of the fruit. A decision rule learner was used to extract simple selection instructions. After delivering these instructions, the data mining process was over.

A second aim may be the definition of a knowledge extraction process. Imagine a marketing department of a large company, interested in maximizing their marketing success. Marketing success is formulated as:

$S_{campaign} = f(A*B/C)$

With A = Number of clients reacting on the campaign.
B = Average benefit per client.
C = Cost of the campaign.

By analyzing the relations between clients and responses to marketing campaigns, a company can optimize their marketing strategies. Assuming different products to be interesting for different markets, there is no one optimal strategy. Also assuming new marketing goals to appear over time, the procedure of obtaining successful marketing campaigns becomes more important than the answer that client X is a good client for product Y.
The business tasks detecting, modeling, predicting, matching and adapting from Part 3 fall into this category.
A variant on this latter aim is the need to track an evolving concept. Here, the problem is how to keep the knowledge model up to date with the underlying distribution. The choice is between learning incrementally (adapting existing knowledge on the basis of new information, thus making the influence of older facts to gradually fade away in the current formulation of the knowledge) and relearning (relearning the knowledge from scratch each time relearning is indicated according to a relearning criterion).
The borders between the learning modes are not always clear. In learning an evolving concept of the knowledge extraction process is of ultimate importance. And relearning a model to adapt to an evolving process can be considered as reformulating knowledge. However, it is important for the success of an application to consider the required application mode, and to design the learning approach to optimally fit the requirements.

References

- Aha, D., D. Kibler. M. Albert. (1991). Instance-Based Learning Algorithms. Machine Learning **6**:37-66
- Brachman, R.J., T. Anand. (1996).The Process of Knowledge Discovery in Databases. In: U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy. (eds.). Advances in Knowledge Discovery and Data Mining. AAAI Press/MIT Press
- Brazdil, P., J. Gama, B. Henery. (1994). Characterizing The Applicability of Classification Algorithms Using Meta-Level Learning. Proceedings of ECML 94. pp83-102
- Engels, R., C. Theusinger. (1998). Using A Data Metric For Offering Preprocessing Advice For Data Mining Applications. In: H. Prade. (ed.). Proceedings of ECAI 1998. Wiley & Sons. pp430-434

- Fayyad, U.M., G. Piatetsky-Shapiro, P. Smyth. (1996). From Data Mining to Knowledge Discovery: An Overview. In: U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy. (eds.). Advances in Knowledge Discovery and Data Mining. AAAI-Press. pp1-37
- Goldberg, D.E. (1989). Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Reading, Ma
- Hunter, L. (1993). Planning to Learn about Protein Structure. In: L. Hunter. (ed.). Artificial Intelligence & Molecular Biology. AAAI/MIT Press
- Kolodner, J. (1993). Case-Based Reasoning. Morgan Kaufmann, San Mateo, CA
- Lim, T.S., W.Y. Loh, Y.S. Shih. (2000). A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms. Machine Learning **40** (3):203-228
- Liu, H., H. Motada. (1998). Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers, Norwell, Ma
- Michie, D., D.J. Spiegelhalter, C.C. Taylor. (eds.). (1994). Machine Learning, Neural and Statistical Classification. Ellis Horwood. http://www.amsta.leeds.ac.uk/~charles/statlog/indexdos.html
- Quinlan, J.R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA
- Rumelhart, D.E., G.E. Hinton, R.J. Williams. (1986). Learning Internal Representations by Error Propagation. In: D.E. Rumelhart, J.L. McClelland, the PDP Research Group. Parallel Distributed Processing. MIT Press, Cambridge, Ma. pp318-362
- Verdenius, F. (1996). Managing Product Inherent Variance During Treatment. Computers And Electronics In Agriculture **15**:245-265
- Verdenius, F. (1999). Entropy Behavior for Selection of Machine Learning Techniques. In: H. Blockeel, L. DeHaspe, Proceedings of Benelearn 1999. pp113-120

6.1.6 Technical integration of data mining

Damiaan Zwietering[7], *Rudi van Lent*[8]

This article looks at technical issues in data mining integration into warehousing environments. This is obviously only part of the larger challenge of integrating data mining into an organization. We show why a database integration approach is a feasible alternative to current developments in application integration. We present recent developments in the field of data mining that solve many of the current integration issues. These developments allow data mining functionality to be integrated into database engines by integrating models into databases. Apart from reducing the complexity of the on-line use of models, this approach also opens the way for fully automated model updates and maintenance of the analysis environment.

The data environment

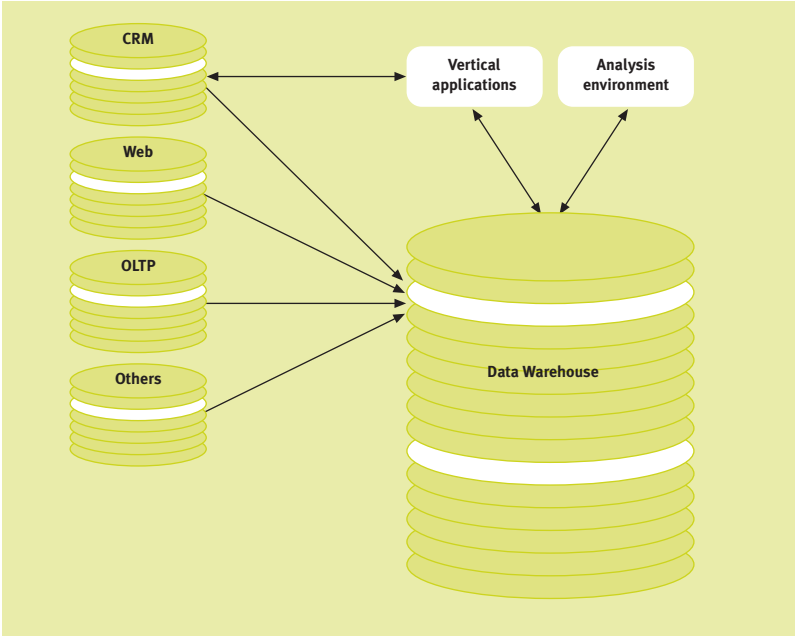


Figure 1
Data environment.

Figure 1 shows a typical data environment. The data warehouse contains data from several source systems to provide an integrated view on several subject areas (e.g. customer or product). Vertical applications directly support activities such as campaign management, using any available data and directly interacting with the CRM environment. This is the domain of the business users. The same data is used by the analysis environment. This is where data analysts try to model the available data and enrich it with new insights for use in vertical applications.

7 Drs Ing D.L.T. Zwietering, zwietering@nl.ibm.com, IBM Global Services, Uithoorn, The Netherlands

8 Ing R.C. van Lent, MSc, rudi_van_lent@nl.ibm.com, IBM Global Services, Uithoorn, The Netherlands

Model deployment

The most important technical aspect of data mining integration is the consistent operational application of models created in an analytical environment. In typical client/server IT architectures, data mining functionality has subsequently been positioned anywhere from server to client. While the server always has had the necessary processing power and detail data available, developments did initially focus mainly on application aspects at the client. This resulted in many desktop query or analysis tools incorporating some form of data mining. However, models developed locally are often hard to deploy to other applications, even without questioning their feasibility, being dependent on locally available data and processing power.

In order to find a solution to this deployment problem, tool vendors started implementing the functionality to export models in C or proprietary macro languages. This may be part of the solution, but model maintenance over several platforms is not trivial, if the only way to apply a model from an application is to recompile a function module. Another solution for the deployment problem is to create mining models in the application environment, avoiding the need for deployment. Data mining functionality is then incorporated in vertical applications.

Functions and skills

The problem with this approach is that the responsibility for model building is shifting from data analysts to decision makers, but the specific skills needed to build reliable models are usually not part of their expertise. A big challenge for data mining on the technical side is data quality and data preparation. These aspects often have such deep technical implications that decision makers typically should not handle them. These problems in the data are the domain of data analysts and as such it is they who should also be responsible for building the various models. However, building relevant models evidently needs the business expertise of the decision maker.
Decision makers would be able to build models, if all data preparation steps could be automated. This however is:

- hardly possible due to the complexity and diversity of data preparation;
- not desirable, because the data preparation stage in itself usually shows many aspects of the data, the organization and its subjects.

The one step that could easily be automated is updating a model to reflect updates in the data. However, if you wanted to automate this step, it would be trivial to automate the whole process just by scheduling the update or base it on triggers. Manual intervention of decision makers would not be necessary.
A functional separation between environments should not cause a physical sep-

aration. A difference in available data between the analysis environment and the application environment may cause a mapping problem. The mapping problem shows up, when a model is based on data or functionality that is not available in the application environment, so a way must be found to map or impute the data that is available.

Application integration

Integration of basic data mining functionality into vertical applications does not solve current issues with deployment:

- Vertical applications often do not have the amount of data available that is needed to build reliable mining models.
- They do not support the required functionality.
- They often run on platforms such as desktop PCs that lack the necessary resources.

When studying the deployment problem, it seems to be an application integration problem. If a vertical application would be able to call an analysis application to apply a model, we would not have to deploy models, but we could simply call on them. At the moment several vendors spend much effort on trying to integrate analysis and application environments in this way.

Models and results

We will take one step back, and look again at what the problem we try to solve exactly entails. Because during the building phase of the model the decision maker has been exhaustively involved, when being applied the internals of the model are no longer interesting. The decision maker is interested in results, not in models. Sophisticated business applications use relational databases as their central data repository or reference point. As such it would be ideal to use the existing mechanism for accessing this data (SQL) to apply mining models. This means that the model should be located close to the data instead of inside either the analysis environment or the application environment. You should be able to 'start' the model from within your application and get the results back in that same application.

Getting results back from a database is trivial for an application with a database connection. Database results are the outcome of a database query, so what we actually need is a way to 'hide' a model in a query. The available mechanism for hiding functionality in a query is a database function, so the model could become a database function. A disadvantage of this approach is that a deployment problem crops up again, because of the maintenance needed on this database function. The most flexible solution would be a generic database function with the model to be applied as a parameter of that function. The only question remaining is where to store models.

Storing models

To keep away from the deployment problem, the analysis environment should be able to store models where it is easy for a database function to access them. Such a generic way of storing models would also enable applications to transparently access the results from applying the model. Recent developments have resulted in a generic language to describe data mining models in a format that can be stored as records in a database table. We can now create a table holding the different models, in such a way that applying the model on the data becomes a database join between a table containing models and various tables containing data. When model application becomes a database join, it can be hidden by a database view. A SQL query containing a model name and a data source will simply produce records with the model's results. In other words, every application that can issue SQL is able to transparently use dependable data mining models.

With this solution data mining also becomes an integral part of all the capabilities we recognize in current relational database systems. Examples are:

- Backup and recovery of data mining models along with data.
- Abstraction through database views (e.g. automatic selection of the most recent or best model).
- Security, scalability, parallel processing, hardware clustering, etc.

By using this approach, technical model deployment issues are reduced to simple inserts, updates or deletes on database tables containing models.

PMML

An independent association (Data Mining Group, DMG[9]) of data mining vendors has specified an open standard based on XML to describe data mining models (Predictive Model Markup Language, PMML). Also, generic application functions have been defined for several database management systems as extensions of SQL.
A typical example of such a design is shown in Figure 1.

This design naturally supports functional separation of analysis and application environments, while maintaining logical integrity. Models can be created in several analytical applications, based on the same data that is available to the application environment. These models are stored as PMML in one or more database tables. Any application that accesses data can call a single database function to apply any chosen model on this data.

9 http://www.dmg.org

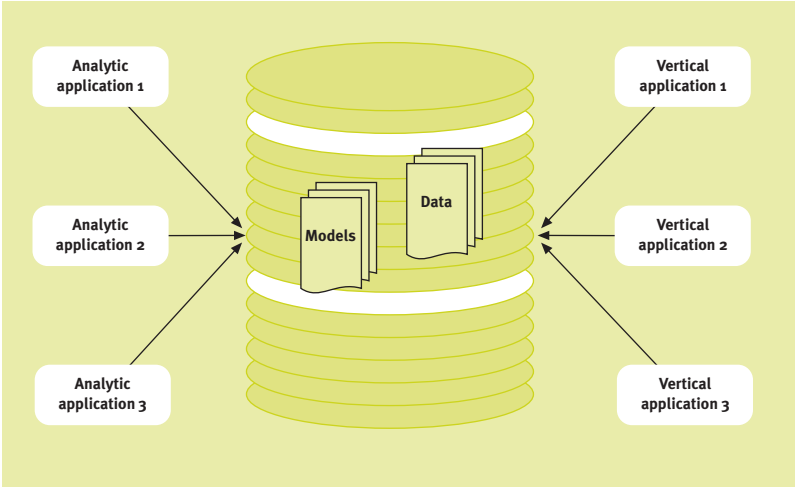


Figure 1
Model building and application environments.

Current developments

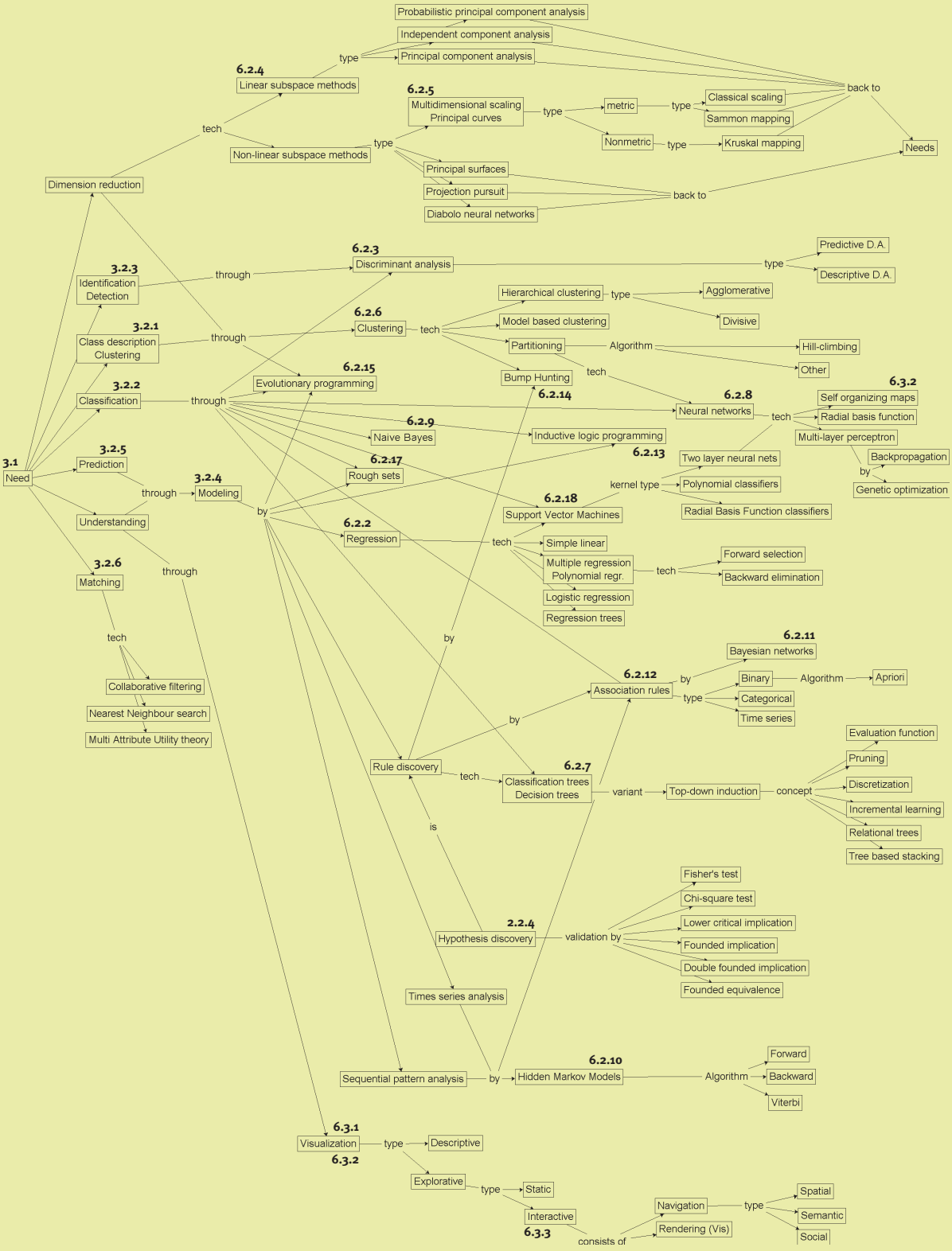
We have seen that current data mining architectures enable both dedicated analysis environments and transparent application environments. At the same time, data warehousing is moving from batch updates to on-line data processing. So, what we need is not just an easy way of making models available, but also an easy way of updating models, when new data arrives. The hard work of creating models is mainly in the first creation, so automated periodic updates can already be provided without analyst intervention. The current challenge is in finding ways of automatically detecting the necessity for updating models or creating new models based on the availability of new data.

6

6.2 Techniques

This chapter discusses a selection of techniques which can be used in the analysis phase of the KDD process. The techniques discussed here were selected by the participants of the project because of their present or future relevance for data mining. Many other techniques exist, and many variations of techniques for specific purposes. In the sections of this chapter, the authors focus on general principles and common applications of the techniques. Figure 1 gives an overview of the techniques discussed in this publication (mainly in this chapter), related to the basic business tasks or needs as discussed in Chapter 3.1. These tasks are in turn illustrated with cases in Chapter 3.2.

Figure 1 (opposite page)
Overview of the data mining technology from Chapter 6.2 (See also Chapter 3.1) in a need-based tree structure.


3.1
Need
3.2.1
Class description
Clustering
3.2.2
Classification
3.2.3
Identification
Detection
3.2.4
Modeling
3.2.5
Prediction
3.2.6
Matching
Understanding
through
Dimension reduction
tech
6.2.4
Linear subspace methods
Non-linear subspace methods
type
Probabilistic principal component analysis
Independent component analysis
Principal component analysis
6.2.5
Multidimensional scaling
Principal curves
Principal surfaces
Projection pursuit
Diabolo neural networks
metric
Nonmetric
Classical scaling
Sammon mapping
Kruskal mapping
back to
Needs
6.2.3
Discriminant analysis
Predictive D.A.
Descriptive D.A.
6.2.6
Clustering
Hierarchical clustering
Agglomerative
Divisive
Model based clustering
Partitioning
Algorithm
Hill-climbing
Other
Bump Hunting
6.2.15
Evolutionary programming
6.2.14
6.2.8
Neural networks
6.3.2
Self organizing maps
Radial basis function
Multi-layer perceptron
Backpropagation
Genetic optimization
6.2.9
Naive Bayes
Inductive logic programming
6.2.13
Two layer neural nets
Polynomial classifiers
Radial Basis Function classifiers
kernel type
6.2.17
Rough sets
6.2.18
Support Vector Machines
6.2.2
Regression
Simple linear
Multiple regression
Polynomial regr.
Logistic regression
Regression trees
Forward selection
Backward elimination
by
6.2.11
Bayesian networks
6.2.12
Association rules
Binary
Categorical
Time series
Apriori
Collaborative filtering
Nearest Neighbour search
Multi Attribute Utility theory
Rule discovery
6.2.7
Classification trees
Decision trees
variant
Top-down induction
concept
Evaluation function
Pruning
Discretization
Incremental learning
Relational trees
Tree based stacking
is
2.2.4
Hypothesis discovery
validation by
Fisher's test
Chi-square test
Lower critical implication
Founded implication
Double founded implication
Founded equivalence
Times series analysis
Sequential pattern analysis
6.2.10
Hidden Markov Models
Forward
Backward
Viterbi
6.3.1
Visualization
Descriptive
6.3.2
Explorative
Static
Interactive
6.3.3
consists of
Navigation
Rendering (Vis)
Spatial
Semantic
Social

6.2.1 Basics and terminology

Jeroen Meij

In this section we will discuss some basic elements and conventions from the wide field of knowledge discovery, logic and statistics. It is a very condensed introduction, emphasizing conventions and notation.

Population and sample

In the statistical context, the term population is used for the total collection of items or individuals we are investigating. However, since the population generally is too large to collect all variables from, the variables are collected for a much smaller group, randomly selected from the population. This sample is analyzed. When the sample is large enough, we can draw conclusions about the population from the results of the sample analysis.
The difference between the value of a property of the population and the value of the corresponding property of a sample is referred to as sampling error. It is a feature of the random sampling process itself, and some degree of sampling error cannot be avoided in statistical experiments involving real populations.

Variable

A variable can be seen as a specific property of each of a population of things. For example, in reference to a population of human beings, a person's height, weight, age, annual income, years of school completed, hair color, favorite flavor of ice cream, favorite political candidate, etc. are all variables.

In statistics, we often wish to characterize relationships between variables (for example, we might wish to determine whether the value of a person's annual income is related to the value of their car color, or perhaps, whether the percent protein in wheat kernels is related to how much fertilizer was applied to the field during growth, etc.). In cases such as these, we still retain the distinction between independent variables and dependent variables. The value of the dependent variable is thought to be determined, or at least influenced, by the values assigned or observed for the independent variables. When the value a variable is given results from a random process (a process in which the specific result is not predictable with certainty in advance), we refer to it as a random variable. Random variables play a large role in statistical work, since we are mostly concerned with properties of members of random samples.

Measurement scales

(With kind permission of [Statsoft, 1999]). Variables differ in 'how well' they can be measured, i.e. in how much measurable information their measurement scale can provide. There is obviously some measurement error involved in every measurement, which determines the 'amount of information' that we can obtain. Another factor that determines the amount of information that can be provided by a variable is its 'type of measurement scale'. Specifically variables are classified as (a) nominal, (b) ordinal, (c) interval or (d) ratio.

- Nominal variables allow for only qualitative classification. That is, they can be measured only in terms of whether the individual items belong to some distinctively different categories, but we cannot quantify or even rank order those categories. For example, all we can say is that 2 individuals are different in terms of variable A (e.g. they are of different race), but we cannot say which one 'has more' of the quality represented by the variable. Typical examples of nominal variables are gender, race, color, city, etc.
- Ordinal variables allow us to rank the items we measure in terms of which has less and which has more of the quality represented by the variable, but still they do not allow us to say 'how much more'. A typical example of an ordinal variable is the socioeconomic status of families. For example, we know that upper-middle is higher than middle, but we cannot say that it is, for example, 18% higher. Also this very distinction between nominal, ordinal, and interval scales itself represents a good example of an ordinal variable. For example, we can say that nominal measurement provides less information than ordinal measurement, but we cannot say 'how much less' or how this difference compares to the difference between ordinal and interval scales.
- Interval variables allow us not only to rank order the items that are measured, but also to quantify and compare the sizes of differences between them. For example, temperature, as measured in degrees Fahrenheit or Celsius, constitutes an interval scale. We can say that a temperature of 40 degrees is higher than a temperature of 30 degrees, and that an increase from 20 to 40 degrees is twice as much as an increase from 30 to 40 degrees.
- Ratio variables are very similar to interval variables; in addition to all the properties of interval variables, they feature an identifiable absolute zero point, thus they allow for statements such as x is two times more than y. Typical examples of ratio scales are measures of time or space. For example, as the Kelvin temperature scale is a ratio scale, not only can we say that a temperature of 200 degrees is higher than one of 100 degrees, we can correctly state that it is twice as high. Interval scales do not have the ratio property. Most statistical data analysis procedures do not distinguish between the interval and ratio properties of the measurement scales.

Null-hypothesis

After being chosen for testing, a hypothesis is called the Null-hypothesis. To determine the validity of this hypothesis, tests are performed. It is common to formulate the Null-hypothesis to indicate that there is no difference or no relation between groups or the variables investigated. A significant relation or difference would mean the rejection of the Null-hypothesis. We can illustrate this with the following example of deterministic hypothesis testing:

Hypothesis: *Two objects of different weights will fall at the same speed.*
Test: Drop two canon balls, one large and one small, from the tower of Pisa.
Outcome: The two canon balls did indeed strike at nearly the same instant.

General structure of this method:

1 According to the hypothesis an observable quantity x (time between impact of the two balls) should have the value x_0 (0 in this case).
2 If we observe a value of x different from x_0, we must reject the hypothesis.
3 The observation that x=x_0 serves to support the hypothesis.

Statistical significance or p-value

The statistical significance is usually expressed in a p-value. This is the probability of a difference or relation found in the sample data, which is not present in the population studied. When the difference in the sample is small, the p-value is large, expressing that it is highly possible that there is no difference in the population. With small p-values there a real difference in the population is likely to exist. For example, a p-value of 0.05 (i.e. 1/20) indicates that there is a 5% probability that a relation between variables is found in our sample, whereas, in fact, the relation is coincidental. Thus, for small p-values (< 0.05) we can reject the Null-hypothesis, which states that there is no difference or relation. Generally, conclusions based on levels of significance, which are 0.05 or 5% are considered acceptable.

Confidence interval

The confidence intervals for specific statistics (e.g. means, or regression lines) give us a range of values around the statistic where the 'true' (population) statistic can be expected to be located with a given level of certainty (the confidence level).
Whenever we estimate some property of a population based on observations for a random sample of that population, we will attach a level of confidence to our estimate. The level of confidence is a number on a scale of 0 to 1, written as a percentage, where numbers near 100% indicate a high likelihood that our estimate is correct as stated. A level of confidence of 100% would mean we are certain our estimate is correct, whereas a level of confidence of 0% means we are

certain our estimate is incorrect. For most work, a level of confidence of 95% is considered acceptable.
Confidence intervals can be given for point estimates and for interval estimates.

Data, information and knowledge

Data set

Usually, data are a sample of the universe around us. This sample may be picked deliberately, with a special purpose in mind, or it may be collected for another purpose than the one we are pursuing. There are also cases in which just as much data as possible is collected — using the sensors available.

This data is stored in a database or data set. This set would ideally be a random sample of the part of the universe we are trying to learn something about, with all the relevant properties or attributes present.

Training set and test set

The data set is often divided into a training set and a test set, especially when we want to extract general rules. The training set is used to extract rules and patterns, which are tested with the data in the test set. General knowledge will be valid in the test set as well.
In a case where we have found rules that correctly describe the training set, but are not valid in the test set, we have 'overfitted' on the training set. In fact we have described it so precisely that the peculiarities of the training set have been integrated in the rules.

Data, information and knowledge

For a definition of the terms, please refer to Section 6.1.2. For our purposes, the most important thing is that the user wants information, not data. This information should be presented in such a way that the user can convert it into knowledge.

General and specific knowledge

Learning from our data can take many forms. An important distinction is general versus specific knowledge.
General knowledge is acquired when we are learning general rules from our data, that will be valid in a part of the universe larger than the sample we have collected.
Specific knowledge is when we just want to learn about our data set, and do not want to apply the rules found to other data. For instance, in mining a text collection or a multimedia collection, you might just want to know what is in the collection, and make the collection more accessible.

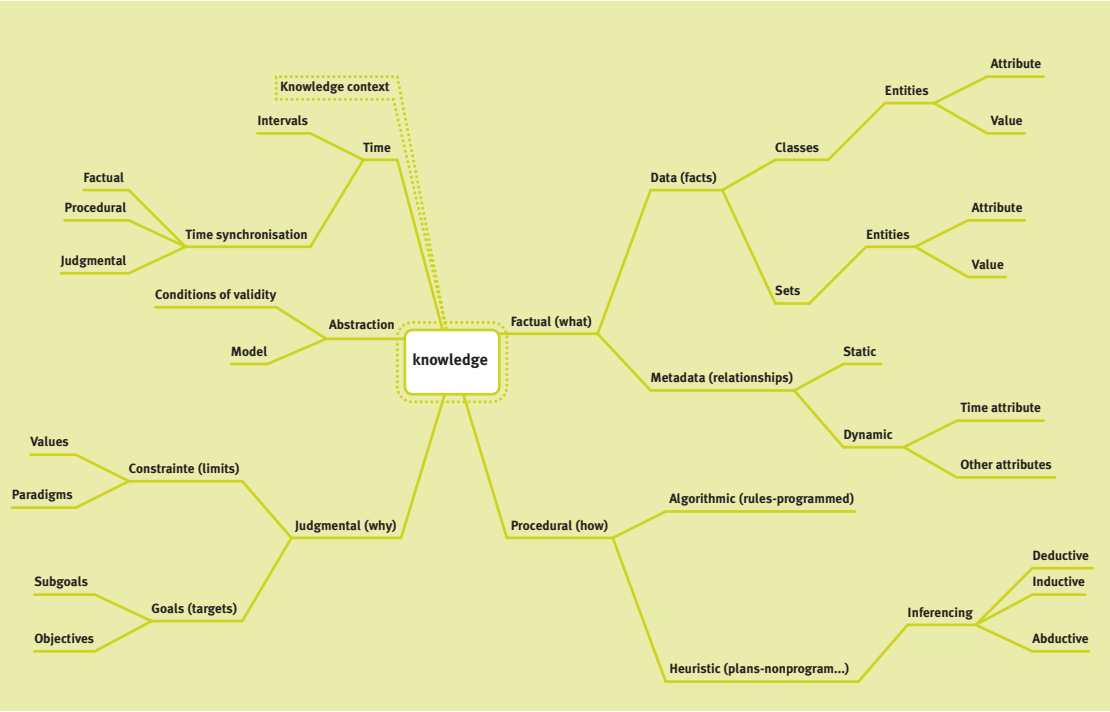


Figure 1
Aspects of knowledge.

Inferring information

Deduction

When we use the data and relations that are explicitly available in a database, we can deduce all kinds of information. However, this information had to be inserted in the database beforehand. For instance, from a (training set) table of a tennis teacher, we can deduce that students John, William and Sandra live in town T. This does not say anything about any of the members of the test set (or the rest of the world), since this rule does not apply (John, William and Sandra are not present in the test set). When performing queries in a database system, information is deduced.

Table 2
Example data set.

Name	Technique	Endurance	Strength	Competed	Prize	Postal code	Address	Town	Phone area	Phone#
John	Y	Y	Y	2001	Y	2514 AP	Tussengracht 23	The Hague	070	3029
William	N	N	Y	2000	N	2513 TS	Middelstraat 45	The Hague	070	2145
Chris	N	Y	Y		N	2500 PP	Dorpsstraat	Rijswijk	070	1632
Peter	Y	N	Y	2001	N	2665 AS	Afrikalaan 234	Zoetermeer	079	1189
Sandra	Y	Y	N	1999	Y	2510 BE	Somestreet 3	The Hague	070	1234
Marian	N	N	Y	2001	N	2553 AS	Derdestraat 12	The Hague	070	4442

Induction

When we are looking to derive general or higher order information from the training set, we try to find regularities. A simple model of the data set is defined, from which general rules are created. Suppose the database of the tennis teacher contains information of his students. Per student, fields are present with items like the student's techniques, endurance, strength, winning a prize, joining the last year's competition, etc.
The 'training' part of the teacher's data set is shown in Table 2.

By induction, we can generate more general rules like 'All people in The Hague have the phone area code 070'. A rule like this could have some validity in the test set and the real world as well. Note that the rule is not necessarily true, but the data in our training set does suggest that it is.

Representation

The information induced from a database can be represented in many ways. The representations can be natural language based, mathematical and graphical and everything in between.
Induction has strong roots in logic and through logic in philosophy. The Greek philosopher Aristotle laid the foundations of logic theory, which was further developed by the Stoics and by scholastic philosophers in the Middle Ages. From the late 19th century, logic has developed continuously.
Defined as the science of formal principles of reasoning or correct inference, logic is strongly intertwined with every effort to convert data to knowledge. Inset 1 describes the three basic principles of Aristotelean logic.

Inset 1: Aristotelean principles

A subject is an individual entity, a man or a house or a town. It may also be a class of entities, for instance all men, all towns with less then 1,000 inhabitants etc.
A predicate is a property or attribute or mode of existence, which a given subject may or may not possess.

Identity
Everything is what it is and acts accordingly.
A* is *A
Non-contradiction
It is impossible for a thing to be and not to be, to belong to and not to belong to.
***A* and non-*A* cannot be the case.**
Either or
Everything must either be or not be. A given predicate either belongs or does not belong to a given subject in a given respect at a given time.
Either *A* or non-*A* [Simpson, 1999]

Propositional logic

The propositional representations stay very close to our natural language. Take the following rule:

If *Town* is *The Hague* then *Phone area is 070*.

More mathematically the propositional statement would be:

Town = The Hague → Phone area = 070.

We say the conditional value for the attribute '*Town*' is *The Hague*.

For many of the words from natural language that have a role in reasoning, logic operators have been described that facilitate logic mathematics. Inset 2 lists some important logic operators. Note that these have a lot in common with Boolean logic operators.
When we focus on the relationships between technique, strength and endurance and (the chance of) having a prize, we can create a rule like this:

If *Technique* is *Good* and *Endurance* is *Good* then *Prize* is *Yes*.

Technique=Good ∧ Endurance = Good → Prize = Yes

Inset 2: Logic operators

a, b, c,…,z **individuals (constants and variables)**

A, B, C,…, Z **predicates (properties)**

When *M* is the predicate 'to be a man', *a* the individual 'Socrates'

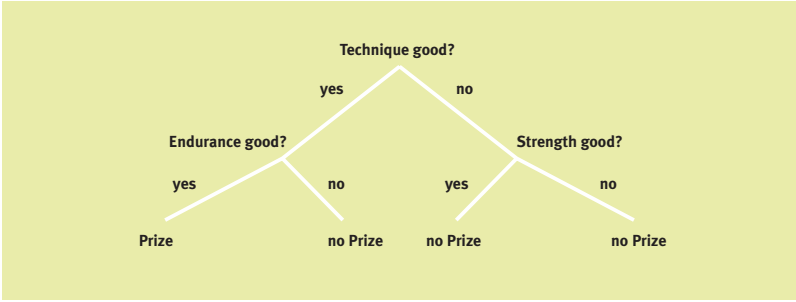
Ma **denotes 'Socrates is a man', *a* is the argument of *M***

∨	**disjunction of two propositions**	**either *A* or *B*, or both**
∧, &	**conjunction, and**	**both *A* and *B***
→, ⊃	**implications, implies**	**if *A* then *B***
⇔, ≡	**bi-conditional equivalent**	***A* if and only if *B***
¬, ~	**negation, is not**	**it is not the case that *B***
∀	**universal quantifier**	**for all *x***
∃	**existential quantifier**	**there exists *x* such that**

[Simpson, 1999]

A decision tree as in Figure 2 is a more graphical form of a propositional representation. A decision is made on every node, leading to the conclusions at the end of the tree branches.

Figure 2
A simple decision tree.



First order logic

A drawback of propositional logic is the multitude of statements that have to be used, when we want to define a relationship between attribute values within the conditions.
For instance, if we want to describe any of the conditions where the strength is equally good or bad as the endurance of a player, we would have to state:
(Strength = Good ∧ Endurance = Good) ∨ *(Strength = not Good ∧ Endurance = not Good)* → .....

We need first order logic to be able to say something like:
When over the domain *X* the strength value is equal to the endurance value then...
As a first order logical statement:
∀ *X*. (*Strength*(*X*) = *Endurance*(*X*) → ...)
or: for all x when *Strength* equals *Endurance* then...
This has great advantages, when the values for the attributes are not limited to two (here: *Good*/*not Good*). For example, when the values are derived from measurements, they could be represented in a range from 0 to 100, meaning a very long line of propositional conditions.

Other representations

Semantic nets will be discussed in Chapter 5.4, Text mining, and neural nets will be discussed in Section 6.2.8, Neural networks for data mining.

Types of rules

Definite rules

A definite rule is a rule that is correct (true) for all records in the training data set. The rule *Town = The Hague → Phone area = 070*
Is true for all records in the data set of the tennis teacher.

The general form for a definite rule is

$(A_1=a_1) \wedge (A_2=a_2) \wedge \ldots \wedge (A_n=a_n) \rightarrow (D=d_i)$

Where A_i is condition attribute number *i*, a_i is a value within the domain of A_i, D is the decision attribute and d_i is a value in the domain of D, a classification.

Or in words:
Whenever attribute A_1 equals a_1 AND attribute A_2 equals a_2 AND... AND attribute A_n equals a_n is true for an object, then the Class *D* of the object is d_i.

Probabilistic rules

If a rule is not true for all records, but has a certain probability of being correct, it is called a probabilistic rule.
For instance, the rule

Phone area = 070 → Town = The Hague

Is correct in most cases of the training set, but not all: Chris in Rijswijk also has phone area code 070.
If a person has a phone area code 070, then there is a probability of 0.8 that he or she lives in The Hague.

Phone area = 070 → Town = The Hague p=0.8

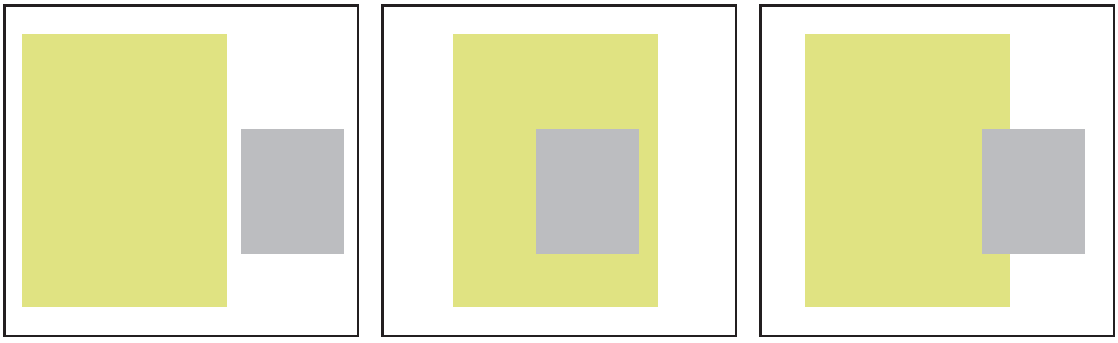
This would be the way to indicate the probability of the rule. Probabilistic rules are also called default rules [Solheim,1996].

Probability conventions

A probability of an attribute (event) *A* occurring is noted as *P(A)*.
There are several ways to calculate with statistically independent probabilities.

A disjunction of *P(A)* and *P(B)*, Either *A* or *B* or Both *A* and *B*
$P(A) \vee P(B) = P(A) + P(B) - P(A \wedge B)$

Figure 3
The probability of a coin falling on the green or the gray surface is equal to the sum of the separate probabilities for green and gray minus the probability of the object falling on the overlap between both surfaces.



Supposing A and B are statistically independent, the probability of the conjunction of A and B, both *A* and *B* is calculated through:
$P(A \wedge B) = P(A)P(B)$

$P(A \wedge B)$ is also shorted to $P(A,B)$

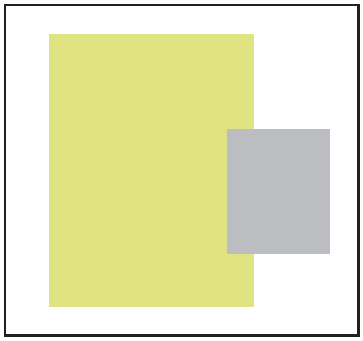


Figure 4
The probability of a coin landing on the overlapping green and gray surfaces equals the product of the separate probabilities for landing on green or gray.

A conditional probability P(A|B) describes the probability for A (or our belief in A) when B is known with absolute certainty.

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

When A and B are fully independent variables

$$P(A|B) = P(A)$$

Our belief in A remains unchanged, whether we know the value of B or not.

Bayes' rule

In the Bayesian point of view, we are not just talking of joint events, but regard condition B as a context pointer, that specifies a context from which A is regarded. A|B stands for an event A given the context B. This leads to the following formulas (Bayes' theorem):

$$P(A \wedge B) = P(A|B)P(B)$$

$$P(A \wedge B) = P(B|A)P(A)$$

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

The probability of an event A can be computed by regarding it from any set of exhaustive and mutually exclusive events B_i, with i = 1,2,...,n and summing the probabilities.

$$P(A) = \sum_i P(A \mid B_i) P(B_i)$$

Or in words, the belief in any event A is a weighted sum over the beliefs in all the distinct ways that A might be realized. [Pearl, 2000; Suber, 1996; Suber, 1999; Gasser, 1997].

Association rules

Suppose A being a set of products, B another set of products (disjoint). A special form of a probabilistic rule is an association rule. An association rule describes the probability that all products in itemset *B* are bought, given that all products in itemset *A* are bought.
An association rule is written in the form:

$A \rightarrow B \mid (c,s)$

c is called the confidence of the rule, the percentage of records with all attributes in *B* having a value true within records with all attributes in *A* with a value true. This is a measure for the strength of the rule. How often does *B* occur when *A* occurs?
s is the support of the rule, the percentage of records that have all attributes in A ∪ B with value true. This is a measure for the statistical significance of the rule. How often does the co-occurrence occur within the total sample?
An example:

{*Peanuts, Crisps*} → {*Soft_drink*} | (95, 60)

95% of the records containing (customers buying) peanuts and crisps also buy soft drinks, while 60% of all records (customers) buy peanuts, crisps and soft drinks.

In general we are interested in all rules that exceed a user specified support and exceed a user specified confidence level.
First, all itemsets F are generated that have a support greater then the set threshold level *s*. These itemsets are called frequent itemsets. For all frequent itemsets, all rules are generated that have a higher confidence level than threshold *c* (note: this is a different confidence than the confidence interval described earlier).

Exploratory and hypothesis driven analysis

Knowledge discovery in data can take on many forms. An essential division is the one between exploratory analysis or discovery and hypothesis driven analysis or verification. Exploratory analysis starts with the data and tries to learn from that data, making few or no presumptions on what to expect. Hypothesis driven analysis starts with a hypothesis, for which data is collected (or selected) and tries to prove or reject the hypothesis. Traditionally, the hypothesis can be based on prior knowledge or intuition.

Exploring data

Exploratory analysis can be useful to develop a 'feeling' about the data set, to determine the most interesting dimensions or variables and to generate hypotheses. These generated hypotheses are also called rules. Many techniques are used for exploring, from visualization to the induction of (large sets of) rules. After hypothesis generation, some sort of hypothesis testing is usually carried out.

Hypothesis testing

The generated or predefined hypothesis is tested by comparing the values in the data set with the values derived from the hypothesis. From this comparison, statements can be made about the quality of the hypothesis (for the given data set). One of the measures for this hypothesis quality is the significance. Others are the confidence and support that have been described earlier. In Section 2.2.4, Mining for scientific hypotheses many more hypothesis tests are described.

References

- Boyle, R., S. Hanlon. (2001). The University of Leeds' Course on Hidden Markov Models. University of Leeds, United Kingdom. http://www.scs.leeds.ac.uk/scs-only/teaching-materials/HiddenMarkovModels/html_dev/main.html
- Schreiber, A.Th. (et al.). (1998). Knowledge Engineering and Management, The CommonKADS Methodology. Version 1.1. The University of Amsterdam
- Simpson, S.G. (1999). Logic and Mathematics. Department of Mathematics. Pennsylvania State University. www.math.psu.edu/simpson/
- Solheim, H.G., Ø.T. Aasheim. (1996). Rough Sets as a Framework for Data Mining. Knowledge Systems Group, Faculty of Computer Systems and Telematics, The Norwegian University of Science and Technology, Trondheim. Project Report. Rough_sets_datamining\index.html
- StatSoft. (1999). Electronic Statistics Textbook. StatSoft, Tulsa, Oklahoma. Electronic Textbook on Statistics. See CD-rom. http://www.statsoft.com/textbook/stathome.html

– Pearl, J., S. Sussel. (2000). Bayesian Networks. UCLA Cognitive Systems Laboratory. Technical Report (R-277). To Appear in: M. Arbib. (ed.). (2001). Handbook of Brain Theory and Neural Networks. MIT Press
– Suber, P. (1996). Symbolic Logic. Philosophy Department. Earlham College. http://www.earlham.edu\~peters\courses\log\loghome.htm. Course Handout
– Suber, P. (1999). Logical Systems. Philosophy Department. Earlham College. http://www.earlham.edu\~peters\courses\logsys\lshome.htm. Course Handout
– Gasser. (1996). http://www.indiana.edu/~gasser/Q351/uncertainty.html. The Trustees of Indiana University

6.2.2 Regression analysis

Dick Bezemer[1]

Introduction

The word regression, to the opposite of progression, has an essentially negative meaning: going back, getting worse. Still, in mathematics and especially statistics regression analysis has come to be a very widely used technique to relate variables to each other. More specifically: for one variable, the dependent one, a prediction or explanation is sought within the relation with one or more 'independent' variables. One tries to go back from independent variable(s) to the dependent one.

Applications are numerous. For instance, the quality of a product is related to manufacturing parameters, income is related to personal characteristics, well-being to health status, development of disease to risk factors. Terminology is sometimes disturbing. Independent variables are also called explanatory variables, covariates, predictors, factors, determinants (in medicine). The dependent variable is also known as response or outcome. The word (in)dependent is of course misleading and even a contradictio in terminis. It only reflects the direction of the relation. However, I will persist in using the word for convenience.

Regression relations should be distinguished from the well-known correlations. Although correlation coefficients are used in regression analysis, conceptually there is a difference. Regression implies direction (from independent to dependent), where correlation only measures the strength of relations. This distinction is very important for applications.

Primarily, regression analysis is a tool for probabilistic prediction. On the basis of data, a regression relation is sought which is descriptive in nature and is used to predict (with uncertainty). The data are converted into knowledge which is, first and foremost, instrumental and probabilistic. Sometimes a regression relation may be used to 'explain', or, rather, suggest to explain, the dependent by the independent(s). But this requires a very considered design in the data gathering. Causal interpretation is still more hazardous, and usually only possible in the context of experimental conditions.

According to the number of independents, the form of the relation and the scale of the dependent, several types of regression analysis arise with always the same basics. In the next sections I will briefly discuss the following types, with examples mainly from the medical field. In a closing section the powers and

1 Dr Ir P.D. Bezemer, PD.Bezemer.Biostat@med.vu.nl, The Vrije Universiteit Amsterdam, Faculty of Medicine, Department of Clinical Epidemiology and Biostatics, Amsterdam, The Netherlands

restrictions of regression analysis will be summarized and some closely related topics will be mentioned.

- Simple linear regression. One numerical (quantitative, interval or ratio) dependent variable is linearly related to one numerical independent variable. This is the basic type of regression analysis.
- Multiple linear regression. One numerical dependent variable is linearly related to several independent variables, which may be numerical, ordinal, nominal or even binary (dichotomous, i.e. nominal with two categories). A special problem in this type of regression analysis is the search for an optimal set of independents. Polynomial regression also falls into this category.
- Logistic regression. One binary dependent variable is related to one (simple) or more (multiple) independent variables, which again may have different scales. The dependent variable is transformed logistically, hence the name.

Simple linear regression

Blood pressure generally increases with age. To describe this relation quantitatively, measurements on the blood pressure may be done for persons of several ages. The results of this, provided that the persons are chosen well, should show the 'normal' or expected blood pressure to be found in persons of a certain age. It is evident that in this example the relation has a direction: from age (independent) to blood pressure (dependent), and not the other way round. I will use this example to discuss simple linear regression.

Suppose we measure the diastolic blood pressure (DBP, the lower one) and assume the relation with age (A) is linear (which is approximately true). Figure 1 shows the measurements (a scatter diagram) and the associated regression line, with DBP=y (vertical, the dependent) and A=x (horizontal, the independent). The line is computed with the 'least squares principle', meaning that the line should be such that the sum of all squared deviations in vertical direction is minimal. The general formula for the regression line is:

$$y = a + bx$$

Most important is the quantity b, the regressions coefficient or slope, representing the change in y per unit change of x. In the example b = 0.25 mmHg/yr, meaning that the DBP increases on average 0.25 mmHg each year. The other quantity, the 'intercept' a, is less important. It represents the value of y when x=0, in the example the theoretical DBP at birth. Once the line is computed from representative data, i.e. a and b, for each x the associated value of y may be derived from it. This value, however, is a mean value, or an expected or predicted value. As is clear from Figure 1, individual data differ from the line. The deviations in vertical direction are called residuals. For example, a 50 year old person

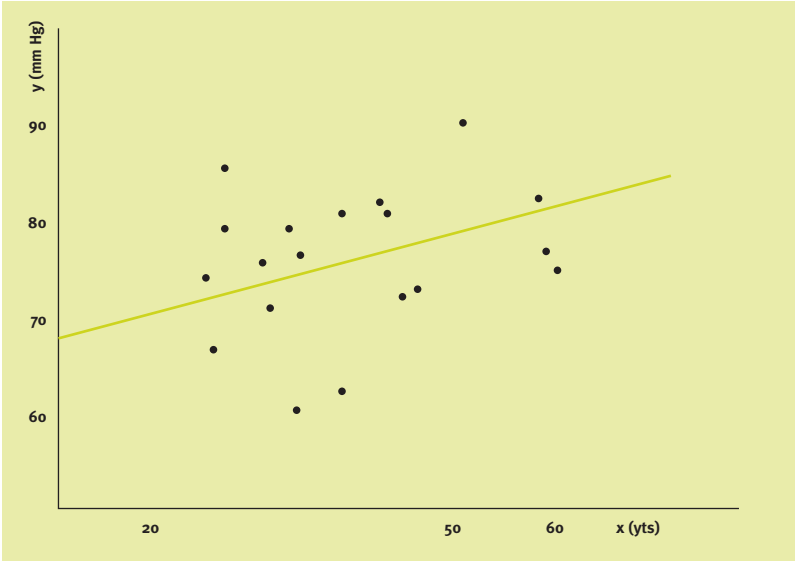


Figure 1
Figure 1 shows the measurements (a scatter diagram) and the associated regression line, with DBP=y (vertical, the dependent) and A=x (horizontal, the independent).

may be expected to have a DBP of 80 mmHg, but the deviation or residual may be as large as 15 mmHg. This example makes also clear why deviations are measured in the vertical direction, that of the dependent variable.

Residuals are very useful quantities, because they contain information about the precision (how large might deviations be?) and the validity (is the line the correct one?) of the regression line. To measure the precision, also of predictions, the standard deviation of the residuals may be computed and intervals derived from it. To look at the validity, the pattern of the residuals is important, especially with increasing x. When, for example, the relation is not linear but curved, the residuals will show it. There is a large body of literature about residuals as diagnostic tools.

Statistically, the computed line is an estimate of the 'population line', because the data are a sample from this population and also because of measurement error. Moreover, the line relates the mean of y with x, with individual data scattering around it. As to the line itself, a confidence interval may be computed for the slope ß (i.e. the population value for b), which is standard in all software. When this interval does not contain the value 0, the slope is real and the relation between y and x is called significant. As to the scatter around the line, a prediction interval may be computed for (measured) individual y's at each x (which includes the uncertainty of the line itself). Confidence interval and the much wider prediction interval should not be confused, since they have a completely different function.

Thus, regression analysis is a very useful and generally applicable instrument to relate variables to each other and to predict the one from the other. However, caution is needed in using the results for the following reasons. The regression line is valid only when:

- the data are a random sample from the population of interest;
- the assumption of a straight line is correct (which may be checked by examining the residuals);
- the scatter around the line is homoskedastic, i.e. the same for all x's (again: look at the residuals); when this is not the case 'weighted regression analysis' should be used.

Moreover, the validity of the confidence interval and the prediction interval around the line depends on the distribution of the residuals. The usual computations assume a normal distribution. Of course extra caution is needed, when extrapolating outside the region where data are present.

Multiple linear regression

As a rule several ('independent') variables are potentially related to the dependent one. For example prognostic variables such as age, condition of the patient and also type of surgery together predict the outcome of surgery for a patient. Or several process parameters influence the quality of the resulting product. Multiple regression analysis is the method of choice in such cases. When the dependent variable is numeric (interval, ratio), multiple *linear* regression is the most widely used and often adequate technique. The aim of this method may be, again, prediction or explanation: (1) which of the independent variables and in what combination predict best the dependent one or (2) which of the independent variables are really related to the dependent one, i.e. on their own and without contamination with the other independents.

The formula describing a multiple linear regression relation is a straightforward extension of simple linear regression:

$$y = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots$$

Now there are several independent variables x, each with their own slope or regression coefficient b. The intercept a is the value of the dependent y, when each of the independents is zero and the b's have the same interpretation as earlier. For each of the b's a confidence interval may be computed and a statistical test performed, indicating whether it is useful to maintain the independent variable involved in the equation. Also, y may be predicted for each combination of values of the independents. It is also possible to judge the strength of the relation between y and the independents together, for which the quantity R^2 is

common. This is the square of the correlation coefficient between y and the combined x's, a quantity ranging from 0 to 1. The scale of the dependent variable should be numerical (as with linear regression), but the scale of the independents is free. When it is numerical, the interpretation is as earlier, but when it is ordinal scores are needed and when it is nominal, dummy values need to be defined. I will illustrate this only for a dichotomy, for example sex. In that case the two possibilities may be coded as 0 and 1, for instance x = 0 for woman and x = 1 for men. The interpretation of the corresponding b is then simply the difference between the mean value of y for woman and that for men, all other x's not changing.

A problem in multiple regression is the interrelation of the x's, a problem which reflects reality and can not be neglected. It implies that the importance of a certain independent variable in the regression equation, i.e. the value of the corresponding b, depends on other independents being present or absent. In Table 1 an example is given for the relation between diastolic blood pressure (dependent) and weight and arm width (independents). First the results of both simple regressions are given and in 3 that of the multiple regression (units deleted).

Table 1
An example of the relation between diastolic blood pressure (dependent) and weight and arm width (independents).

	independents	b	p-value	R^2
1	weight	0.30	0.0006	0.11
2	arm width	1.21	0.0014	0.10
3	weight	0.21	0.14	0.12
	arm width	0.49	0.42	

As is seen from Table 1, the simple b's, and also the corresponding p-values, differ greatly from the multiple ones. The reason is that weight and arm width, obviously, are strongly related. In the simple model the p-values clearly indicate significance (generally when $p < 0.05$), which is lost in the multiple model for length and arm width separately. Also R^2, a measure for the total model, only slightly increases. Thus weight alone is also as good as weight and arm width together. The multiple b's may change again, when a third independent is added. This example illustrates that with interrelated x's it is difficult to find the 'right' regression relation. It is also possible that b, and thus the importance of the corresponding x, increases, when a variable is added. This reflects the complex relations encountered in practice. So much caution is needed, especially when the aim is explanation.

In multiple regression the problem is often how to select a useful subset of independent variables from a larger set of potentially important variables. This problem is complicated, especially when there are strong interrelations. Two main methods exist:

1 *Forward selection*. Independents are added one by one on the basis of maximal contribution (criterion: R^2) to the strength of the relation between y and the combined x's present. Also, the added x should have a significant contribution (criterion: p-value). In this process of building a regression model, values may also be eliminated, when no longer useful. The process stops, when no further variables contribute significantly. This method of forward selection is especially suited to finding the minimal set of variables which, combined in the regression equation, predicts the dependent variable optimal.

2 *Backward elimination*. Starting with all potentially important independents, they are eliminated one by one from the regression equation as long as their contribution is not significant. In the resulting regression model each variable present has an 'own' relation (at least partly) with the dependent, apart from the other independents present. Therefore this method is more suited, when the aim is explanation. Of course, in the process of eliminating variables may also be added again, because at that stage they are related significantly with the dependent.

Multiple regression is a technique which requires many choices and thoughtful interpretations. To simplify, it is wise to pre-select the potential independents on the basis of prior knowledge. As a rule of thumb, the number of independents entering the 'regression process' should be no more than 10% of the number of sample units (for example: persons measured). As to the choice of the significance level, when forward selecting or backward eliminating, this may well be higher than the usual 5%. Otherwise important combinations of independents might be missed, because of complex interrelations.

Finally, polynominal regression should be mentioned, because it is also multiple linear regression. The independents are the linear, the quadratic, the cubic, etc. of the same variable, but each is, as such, linearly related to the dependent variable. Therefore the same computing process applies as well.

Logistic regression

Often the dependent variable is of a dichotomous nature. For example a person has a certain disease or not, a product meets certain standards or not, etc. In that case it is the probability or frequency of the one or the other state which is of interest. This probability (p) is related to one or more independent variables in a regression model. As in linear regression, the independents may be numerical, ordinal or also dichotomous. For reasons of scaling the probability p is transformed into the natural logarithm (ln) of (p/(1-p)), which ranges from minus to plus infinity with p ranging from 0 to 1, and results in 0, when p = 0.5.

This is the logistic transform, hence the name logistic regression.
Now ln (p/(1-p)) is linearly related to one or more independents as earlier:

$$y = \ln (p/(1-p)) = a + b_1 x_1 + b_2 x_2 + ...$$

It appears that many relations in practice are adequately expressed in this way. The quantity (p/(1-p)) is called the odds; when p is, for example, 0.75, the odds are 3 (to 1). Some mathematics is needed to show that each b in the model equals the ln of a ratio of odds, namely the odds belonging to x and x+1. Thus, when x changes one unit, y changes ln OR, where OR is the abbreviation of odds ratio. In the context of relating dichotomous variables to each other, the OR is a well-known measure of the strength of the relation (OR=1 when there is no relation), like a correlation coefficient for numerical variables.

Just like linear regression, logistic regression may be simple or multiple, the latter being much more dominant. Again there is the distinction between prediction and explanation and the difficult problem of finding the adequate independents to incorporate into the (multiple) logistic model. In the medical field a predictive logistic model is increasingly applied for prognosis and diagnosis. An example is the prediction of a person suffering from diabetes mellitus, or not, on the basis of several symptoms being present or not. For many data on a large group of persons, known to suffer or not to suffer from diabetes, a multiple logistic model was found with 10 independents (symptoms or risk factors) by means of forward selection. For illustrative purposes part of this result is shown in Table 2.

Table 2
A multiple logistic model of which part of the result is shown.

independent	b	score	OR (95% CI)
frequent thirst	0.69	3	1.99 (1.22-3.25)
age per 5-year increment from 50 years	0.39	2	1.47 (1.27-1.70)

In this table CI means confidence interval; because both lower limits exceed 1, both OR's (and b's) are significant at the 5% level. The b's are translated into scores, with which for a person with certain symptoms and risk factors it easily may be judged whether a test on diabetes in indicated. For example, a person of 60 years with frequent thirst has score 3+2+2=7, and when this score is above a certain cut off, the test is indicated. Note that for age a unit of 5 years is chosen in this model and note also that only 2 of the 10 independents in the model are shown in the table. The example nicely illustrates in what way the results of a logistic regression might be used in practice.

Summary

Regression analysis is a widely used and indeed very useful tool to relate variables to each other. The relation is one-directional, i.e. one 'dependent' variable is related to one (simple) or more (multiple) 'independent' variables, as opposed to correlations which are non-directional. The regression relations found are primarily suited for predictive purposes, causal interpretation is hazardous and requires a very careful gathering of data. The knowledge into which data are converted is descriptive, instrumental.

There are many types of regression analysis. An important distinction is that according to the scale of the dependent: linear regression with a numerical scale (interval, ratio) and logistic regression with a dichotomous scale (binary). Multiple regression poses special problems, because of interrelated independents, as to the choice of the variables in the ultimate model. Residuals are a powerful tool for checking model assumptions.

Finally, some closely related methods are mentioned, all of them in fact regression analysis for a special application. With discriminant analysis (see Section 6.2.3 predefined groups (of data, persons, etc.) are distinguished on the basis of measured characteristics. Time series analysis tries to find special functions to describe time course of a phenomenon. Survival analysis is a recently developed body of methods in the medical field to describe the probability of a certain event (death, birth, complication of disease, etc.) in time, and of the influencing factors. Also cluster analysis (finding groups which are not predefined) may be seen as related to regression analysis.

Further reading

– Draper, N.R., H. Smith. (1981). Applied Regression Analysis. New York, Wiley & Sons
– Kleinbaum, D.G., L.L. Kupper, K.E. Muller, A. Nizam. (1998). Applied Regression Analysis and Other Multivariable Methods. Pacific Grove, Duxbury Press

6.2.3 Discriminant analysis

Carl J Huberty[1]

Introduction

What is a 'discriminant analysis'? The answer to this question varies from one computer package to another, from one general multivariate book to another, from one research application to another, from one university professor to another, from one workshop to another, and from one dictionary to another.
Discriminant analysis was supposedly originated by Sir Ronald A. Fisher (1890-1962) in 1936 for the purpose of determining a rule to classify plants into one of two taxonomic categories (i.e. species). The rule used to make the category assignment was based on a single linear composite of a set of 12 plant characteristics. (It may be pointed out that Fisher actually gave credit to E.S. Martin and M. Barnard, who applied the idea of category assignment prior to 1936.)
The learning process is a supervised process, where the training set consists of pre-classified observations. When we have two categories, like Fisher we may represent each of the two classes A and B by a linear composite in the form of:

$$Y_A = a_A + C_{1A}X_1 + C_{2A}X_2 + \ldots + C_{12A}X_{12}$$
$$Y_B = b_B + C_{1B}X_1 + C_{2B}X_2 + \ldots + C_{12B}X_{12}$$

When the coefficients C have been calculated, Y_A and Y_B are calculated for every new observation, and the observation is assigned to the group for which it yields the higher Y value.
As discussed by [Morrison, 1990], the Fisher two-group classification approach may be extended to the multiple-group context. This approach gets fairly complicated for, say, five groups because of the number of linear composites to be determined. Fisher and others for decades, termed such a linear composite a 'discriminant function'.

So, what view of 'discriminant analysis' did Fisher take? In terms of the apparent purpose of his analysis, one might conclude that the purpose was to develop a classification rule to be used to assign analysis units to one of two criterion groups. Or, one might conclude that the purpose was to determine a linear composite of the response variables that maximally separates/distinguishes the two groups. To me, at least, these are two different purposes.
If the purpose of the analysis is to determine a rule to be used in assigning an analysis unit to one of the two or more groups, then I prefer to call the analysis a predictive discriminant analysis (PDA). If, on the other hand, the purpose of the analysis is to provide a description of (two or more) group differences, then I prefer to call the analysis a descriptive discriminant analysis (DDA).

1 Prof C.J Huberty, chuberty@coe.uga.edu, Educational Psychology, The University of Georgia, Athens, GA, USA

A common data set will now be utilized to illustrate a PDA as well as a DDA. It is realized that the two analyses would not typically be run on a given data set. Generally speaking, a researcher has either a group-prediction problem (calling for a PDA) or a group-difference-description problem (calling for a DDA). But, to simplify the present discussion only one data set will be considered. A description of the data set is given by [Huberty, 1994a]. Briefly, there are k=4 groups of N=442 post-high school students — with group membership determined in 1962 — on each of whom there are 15 response variable measures (seven cognitive measures, five interest measures and three temperament measures) that were obtained in 1960. The four groups are: (1) n_1 = 89 students in a teacher college; (2) n_2 = 75 students in a vocational school; (3) n_3 = 78 students in a business or technical school; and (4) n_4 = 200 students in a university. The two analyses, PDA and DDA, will now be discussed.

To do either analysis, one needs to rely on the use of a computer package. For the analyses discussed herein, the two packages considered are SAS and SPSS. Prior to reporting PDA or DDA results, two types of (preliminary) information should be reported: those pertaining to design and to general results. The latter pertains to data conditions and descriptives, including covariance matrices. These details will not be fully discussed herein, see [Huberty and Hussein, 2000].

Predictive discriminant analysis

From the PDA perspective, the 15 response variables would play the role of 15 predictor variables, and the grouping variable would play the role of the criterion variable (defined by the four groups). The primary purpose of the study, then, would be to determine a rule based on the 15 predictors to predict membership in, or identification with, one of the four criterion groups. There are three different forms of a 'rule'.

- One form is four sets of weights that would be applied to the respective predictors. Each set of weights defines a 'classification function'. A classification function score is determined for each analysis unit for each group; a unit is assigned to the group with which the largest function score is associated.
- The second form is that of a distance from each unit score vector to the centroïd of each group. A unit is assigned to that group which yields the smallest distance.
- The third form involves something called a posterior probability. This is the (estimated) probability that a unit belongs to a group, given the unit's predictor score vector, denoted P($G|\underline{X}$). [SPSS uses the notation P(G|D).] Such a probability may be determined for each unit for each group (the posterior probabilities typically considered are those based on multivariate normality). The unit is assigned to the group with which we associate the largest posterior probability.

It turns out that the three rule forms will yield identical group assignment, provided the same numerical information is used with all three.

The SPSS program used to conduct a PDA is DISCRIMINANT, while the SAS program is DISCRIM. (The current printing format for DISCRIMINANT is different from that found in [Huberty, 1994], while the SAS format is now pretty much the same as that prior to 1994.)

Data conditions

Covariance matrix equality

Assuming approximate multivariate normality of the 15-score vectors in the four corresponding populations, the next analysis step is to assess the equality of the *k* covariance matrices. (With this assumption being reasonable, what will be discussed is a 'normal-based PDA'). The reason covariance matrix equality is examined is to determine the type of rule to use. If it is reasonable to assume equality, then each form of the classification rule is based on the error covariance matrix, as opposed to being based on the separate group covariance matrices. This basis will yield a linear classification function (LCF) for each group. Likewise, the posterior probabilities and the distances will be based on the error covariance matrix. If, on the other hand, it could have been concluded that the covariance matrices were not 'in the same ballpark', then a quadratic classification function (QCF) would be utilized.

The covariance matrix equality assessment may be based on two transformations:

- An *F* transformation of the Box *M* statistic.
- A chi-squared (X^2) transformation of the Box *M* statistic [see Huberty, 1994b].

Both the *F* and X^2 tests are very powerful, the latter more so than the former. So, for starters, one may consider the *P* value associated with the *F* test, supposedly the less powerful of the two tests. Invariably, this *P* value is 'small'. Therefore, what is suggested is to examine the equality of the (natural) logarithms of the *k* covariance matrix determinants and of the error covariance matrix. If the 'eyeball examination' of the *k*+1 logarithms indicates approximate equality — a judgment call — then proceeding assuming matrix equality may be reasonable. Another eyeball examination may be made by looking at the *k*+1 covariance matrix traces. In this context, a trace is a sum of the 15 predictor variances.

For the current data set, $M = 575.56$, $F(360, 205005.8) \doteq 1.4846$, $P \doteq .0000$. The determinant logarithms are: G_1, 47.1; G_2, 48.8; G_3, 46.9; G_4, 50.5; and Error, 50.2. And the five traces are: G_1, 1519.7; G_2, 1434.8; G_3, 1337.2; G_4, 1555.5; and

Error, 1491.2. The five numbers in each set appear to me, at least, to be 'in the same ballpark'. So, even though the *P* value is quite 'small', the other information does not imply matrix inequality.

Prior probability of group membership

There is another very important piece of numerical information to incorporate into a classification rule. This is a prior probability of group membership. In general, there are *k* priors, one for each group. A prior, q_g, reflects an estimate of the probability of unit membership in population *g*. It is prior in the sense that this is the probability of population membership before unit predictor scores are known. In other words, the priors reflect the relative sizes of the populations. The corresponding population priors are sometimes called base rates.
I do not recommend the casual use of equal priors; rather, an attempt should be made to estimate the relative population sizes — 'expert' advice? — as a basis for the priors to be used. Another basis for the priors is the size of each group; that is, some researchers use $q_g = n_g / N$. This basis would be appropriate only if the group sizes reflect the relative sizes of the *k* populations. Such priors may be appropriate if a proportional sampling plan is used. The priors used with current data set are .20, .20, .20, and .40.

Linear classification functions

The predictor variable weights/coefficients for the *k* LCFs are outputted by both SPSS DISCRIMINANT (which uses the inappropriate subheading, 'Fisher's linear discriminant functions'), and SAS DISCRIM (with the command, POOL = YES). With each LCF there is a constant that includes the respective prior [see Huberty, 1994c]. The LCF weights, themselves, are virtually non-informative. They may, however, be useful as a classification rule to be used with new data. The predictor variable weights are mathematically determined in such a way that the hit rates are maximized. The weights for the *g*th LCF are given by

$$L_{ug} = \left[\bar{X}'_g S^{-1}\right]' X_u + \left[-\frac{1}{2}\bar{X}'_g S^{-1} X_g + \ln q_g\right]$$

with u= 1,2,..., n_g and g = 1,2, ...,k.

PDA rule assessment

Hit rates

So, how 'good' a rule do we have? To address this question, we need to estimate the true proportions of hits. A 'hit' occurs when a unit originally identified with a group is assigned to the same group via the use of the defined classification rule. Just as in multiple regression analysis, determining a rule based on a given data set and then applying that rule to the same data set — an 'internal' analysis —

will yield biased group 'hit rates'. Thus, to reduce this bias, one can build a rule on one data set and then apply the rule to another data set — an 'external' analysis — to yield hit rate estimates. The external hit rate estimation approach that I favor is the Leave-One-Out (L-O-O) method. With this method, one unit predictor score vector is held out, a rule is built on the remaining *N* - 1 score vectors, which is then applied to the unit that was held out. Thus, *N* rules need to be developed and each rule is applied to the unit left out. L-O-O results are easily obtained via the use of SAS DISCRIM with the word CROSSVALIDATE in the PROC DISCRIM command. In 1997 the L-O-O option was added to the SPSS version 7.5; thus, using SPSS DISCRIMINANT, linear L-O-O results may be obtained by clicking on 'Leave-One-Out classification'.
The linear L-O-O results for the current data set are reported in Table 1 taken from the SAS DISCRIM output in [Huberty, 1994d].

		Predicted group				
		1	2	3	4	
Actual group	1	30 (33.7)	24	16	19	89
	2	19	25 (33.3)	9	22	75
	3	20	13	12 (15.4)	33	78
	4	17	12	11	160 (80.0)	200
		86	74	48	234	442

Table 1
Linear L-O-O results. Separate-group hit rates are reported in parentheses.

The four group L-O-O hit rate estimates (in percents) are 33.7, 33.3, 15.4, and 80.0. So, it is obvious that the only 'respectable' separate-group hit rate is for Group 4. The total-group hit rate may be obtained in two ways. SPSS uses what I call the 'diagonal method', obtained by simply summing the main diagonal elements and dividing the sum by *N*. For this case, the total-group hit rate is 51.4. SAS uses what I call the 'weighted method', obtained by weighting the separate-group error rates using the group priors. For this case, the total-group hit rate is found as follows:

$$1 - [(30/89).20 + (25/75)(.20) + (12/78)(.20) + (160/200)(.40)] \doteq .515.$$

The reason this hit rate of 51.5 is close to that obtained using the diagonal method (51.4) is because the group sizes pretty closely reflect the four priors used. I judge in most research situations involving a PDA that separate-group hit rates would be of more interest than the total-group hit rate.

How much better than chance?

Next, a question that invariably would be of interest is: How good are the obtained results? A more specific issue related to the obtained *k*+1 hit rates is the 'chance' issue. That is, is the hit rate of interest better than that obtainable by chance? And, if so, how much better than chance? Details of answering these two questions are given by [Huberty, 1994e]. As an example, consider the hit rate for Group 4, 80.0. Now, we could get .40 (200) = 80 hits for Group 4 by chance, while our linear external rule yielded 160 hits.

Thus, $(160 - 80) / \sqrt{80\,(200 - 80) / 200} \doteq 11.547$. This standard normal statistic value yields a very small *P* value. Thus, we may conclude that the observed Group 4 hit rate is better than a chance hit rate. Now we address the question: How much better than chance? To address this question, one may use the following index [Huberty, 1994f]: $I = (H_0 - H_e) / (1 - H_e)$, where H_0 denotes the observed hit rate, and H_e denotes the chance hit rate. For the above results, $I = (160/200 - 80/200) / (1 - 80/200) \doteq 0.67$. That is, about 67% fewer classification errors would be made using the obtained rule than if the classification were done by chance.

What was discussed above is, to repeat, information pertaining to a *linear external* PDA. The descriptor *linear* is used because the *k* composites of the predictor variables are linear — the error covariance matrix is used to determine the composite weights. The descriptor *external* is used, because the PDA rule is based on one data set and then applied to another data set to obtain hit rate estimates — here 'external' refers to L-O-O. Now, if it is concluded that the *k* group covariance matrices are clearly not equal, then the rule to be used would be based on separate group covariance matrices. That is, the predictor variable composites would be quadratic in form — these composites are called quadratic classification functions (QCFs). To obtain legitimate quadratic PDA results, one can not use the SPSS package program. To obtain quadratic external PDA results, the researcher may rely on the SAS DISCRIM program — using the commands POOL=NO (or, POOL=TEST, assuming the chi-squared test will yield a 'small' *P* value) and CROSSVALIDATE.

Hit rates based on a quadratic PDA will usually be higher than linear ones. With the current data, this was not the case as the respective separate-group hit rates are 34.8, 38.7, 20.5, and 68.5, while the diagonal total-group hit rate is 48.2. Inferentially, a problem with quadratic results is that the hit rate estimates are not as precise as linear hit rates.

Outliers

When reviewing the applied research literature in PDA results are reported, it is very common that group results — as discussed above — are the focus. But some relevant information may also be obtained by examining two types of unit

results. One type of unit result that may be informative pertains to the identification of outliers. An outlier would be a unit that is assigned to a particular group, but has a score vector not very close to the centroïd of that group. To quantitatively identify such units, one needs to rely on the use of the SPSS DISCRIMINANT program — when a linear external analysis is being conducted. The program outputs something called a typicality probability, which SPSS denotes by P(*D*>d|*G*=*g*) — it may also be denoted by P($\underline{X}$|*G*), the inverse of a posterior probability, P(*G*|$\underline{X}$). As discussed by [Huberty, 1994g], P($\underline{X}$|*G*) may be thought of as a tail area of a chi-squared distribution, a referent distribution for a particular squared distance. So, an outlier may be identified by a 'small' associated P($\underline{X}$|*G*) value. Four examples of outliers are given in Table 2: units 20, 119, 169, and 438. [A total of 20 outliers were identified with P($\underline{X}$|*G*) ≤ .010.]

Table 2
Some specific PDA results; linear L-O-O results taken from SPSS DISCRIMINANT and SAS DISCRIM output.

		$P(\underline{X}_u \mid G)$ **by Group**			
Unit (G)	$P(\underline{X}_u \mid G)$ *	1	2	3	4
20 (1)	.004	.619	.149	.108	.124
55 (1)	.767	.299	.323	.272	.105
110 (2)	.935	.299	.283	.182	.237
119 (2)	.003	.333	.488	.072	.107
169 (3)	.001	.298	.397	.178	.126
224 (3)	.759	.326	.243	.339	.092
322 (4)	.940	.172	.265	.289	.274
438 (4)	.000	.536	.131	.145	.189

* This is a typicality probability for the group to which the unit is assigned.

It is obvious that Unit 20 should be identified with Group 1 (based on the P(1|$\underline{X}_{20}$) value of .619), but yet that the unit's vector of 15 predictor scores is quite distant from the centroïd of Group 1 (as reflected by the small value of the associated typicality probability, .004). What action might be taken regarding such outliers? Assuming there are no data recording/entry errors, should some or all outliers be eliminated from the data set? One thing that might be considered is initial group membership versus predicted group membership — for example, look at Unit 438, a student who was initially identified with Group 4, but was assigned to Group 1. There were other such types of outliers in this data set. The final decision on what to do with outliers is a judgment call on the part of the researcher(s).

Fence riders

The second type of unit result that may be informative is the identification of in-doubt units, or 'fence riders'. These are units whose predicted group member-

ship is not 'clear'. For example, from Table 2 it is perhaps obvious that group membership for Units 55, 110, 224, and 322 is not that clear. Consider Unit 110 with (at least) two posterior probabilities that are pretty 'close' numerically. That unit was assigned to Group 1, but it may be concluded that this unit could be assigned to Group 2 from which it emanated. There may very well be research situations in which the researchers would impose a restriction such as: Assign a unit to a group only if the assignment probability is 'high'. This may be accomplished via SAS DISCRIM using the command THRESHOLD. For the current data set, a THRESHOLD (posterior probability) value of .35 or .40 may be reasonable. From Table 2, we see that using THRESHOLD = .35 would delete the in-doubt Units 55, 110, 224, and 322. By examining predictor profiles of in-doubt units, some interesting unit characterizations may be revealed.

Predictor deletion

There are two other pieces of information associated with a PDA that may be helpful. One is that of predictor variable deletion. A way of looking at a research situation in which group-membership prediction is of primary interest is to develop the best classification rule using the data on hand. It has been my experience that, given *p* predictors, a better rule may be determined using less than *p* predictors. Now the problem is to find the best subset of predictors — best in the sense that the subset will yield the highest hit rate of interest. That hit rate may be the total-group hit rate, or the hit rate for a particular group. So, the predictor deletion analysis suggested here is the all-possible-subset analysis. Now, with *p* predictors a strict all-possible-subset analysis would call for 2^p-1 analyses. However, there may be some predictors that the research wants to retain, no matter what. That is, it may be desirable to 'force in' some predictors. Whatever, there is at least one computer program that has been written that will do the all-possible-subset analysis[1] (with or without forcing some predictors in).

Predictor ordering

The second piece of information related to a PDA that may be helpful pertains to predictor ordering. That is, which predictor is most important, second most important, ..., and least important. Importance here pertains to contributing to classification accuracy. To accomplish such a predictor ordering, what is suggested is to conduct *p* PDAs, each with *p*-1 predictors. The predictor which when omitted decreases the hit rate of interest the most, would be considered the most important one. [What is very often found is an increase in hit rate as some predictors are deleted.] A set of predictor ranks (with some tied ranks, of course) could then be determined.

The predictor deletion problem and the predictor ordering problem are discussed and illustrated to some extent by [Huberty, 1994h].

1 The author may be contacted regarding the program.

Descriptive discriminant analysis

Descriptive discriminant analysis (DDA) pertains to the analysis of group mean differences with respect to a collection of outcome variables. Many aspects of DDA, as will be seen shortly, serve as a follow-up to a multivariate analysis of variance (MANOVA) or a multivariate analysis of covariance (MANCOVA). As might be surmised, MANOVA is a generalization of univariate analysis of variance (ANOVA) in which a single outcome variable is involved, and MANCOVA is a generalization of univariate analysis of covariance (ANCOVA).

Whereas the design associated with a PDA involves a single grouping variable (or, factor) with two or more levels — which plays the role of a criterion variable — and at least one predictor variable, the design associated with a MANOVA/DDA involves one or more grouping variables — each of which plays the role of an explanatory (or predictor) variable — and at least two outcome/criterion variables. The current discussion will focus on one factor with k levels and p-outcome variables. See [Huberty, 1994i] for discussion related to a two-factor design. As mentioned earlier, the data set used to illustrate the application of a DDA is the same as that used to illustrate the application of a PDA. It should be noted that the role of the set of response variables in DDA (in which they are outcome variables) is reversed from the role played in PDA (in which they are predictor variables).

Data conditions

Just as with the PDA situation discussed above, there are some data conditions that need to be considered prior to conducting the usual MANOVA.

Covariance matrix equality assessment

Multivariate normality of the p-element score vectors in each group/population is, as with PDA, of importance in conducting the statistical test regarding the equality of group covariance matrices. Unless this condition is reasonably met, the usual MANOVA is not applicable. [If it is not reasonable to conclude that the k group covariance matrices are 'in the same ballpark', then a Yao test may be used when $k = 2$ and a Johansen test when $k > 2$ [Huberty & Petoskey, 2000].

The SPSS program to be considered is, again, DISCRIMINANT, while the SAS program is CANDISC — the SAS package also has STEPDISC and GLM that might be considered.

With the current data set, let us proceed as though all data conditions are met, and test the MANOVA null hypothesis that the true four 15-element mean vectors are equal. There are at least four test criteria that could be used [Huberty, 1994j]. The criterion to be considered here is the Wilks lambda which may be transformed to an F test statistic (using CANDISC) or a χ^2 test statistic

(using DISCRIMINANT). For the current data set,
$\wedge \doteq .5696$, $F\,(45, 1260.4) \doteq 5.841$, $P \doteq .0001$, and $\eta^2_{adj} \doteq .327$.
It appears reasonable, to me at least, to conclude that there are group mean-vector differences among the four groups. Proceeding now to a DDA will provide some description and interpretation of the resulting (real?) differences among the four groups.

Linear discriminant functions

What may be obtained using data in a *k*-group, *p*-outcome variable design is a set of linear discriminant functions (LDFs) — SPSS labels them 'canonical discriminant functions'. An LDF is a linear composite of the *p* outcome variables (the weights for the composite are mathematically determined so that the groups are maximally separated with respect to the composite scores). Just as in PDA, 'linear' goes along with the condition of equal covariance matrices. The variable weights are determined via an eigenanalysis [see Huberty, 1994k].

Number of LDFs

The maximum number of LDFs for a given study is the minimum of *p* and *k*-1. For the current data set, there would be a maximum of *k*-1=3 LDFs to consider. Now what has to be determined is the number of LDFs to retain for interpretation/description purposes. This involves a set of *k*-1 statistical tests. If the MANOVA null hypothesis is rejected, it is known by default that at least one LDF should be retained. So then we need to decide if we should retain two or three LDFs. Two more 'tests of dimensionality' are to be conducted. Such test information is outputted by both SPSS DISCRIMINANT and SAS CANDISC. An interpretation of these multiple tests is given by [Huberty, 1994l]. (It should be emphatically pointed out that the significance of individual LDFs is *not* being considered).

Table 3
Tests of LDF dimensions.

Null hypothesis	F	P	Cumulative percent of variance*
No separation on any dimension	5.841	.0001	84.9
Separation on at most one dimension	1.574	.0304	96.8
Separation on at most two dimensions	0.735	.7290	100.0

* The percent of variance in the 15-variable system accounted for by the LDFs (in a cumulative manner).

The test information for the current data set is given in Table 3. It is pretty obvious that we do not need to consider retention of the third LDF in describing the resultant group differences. But we do conclude that there are group differences with respect to two dimensions; that is, two LDFs.

Structure r's

So, our next step is to attempt to 'interpret' the two LDFs. (The LDFs are, computationally speaking, like principal components). To describe the LDFs, then, we look at the 15 correlations — i.e. structure *r*'s — between each outcome variable and each LDF; recall that an LDF is a linear composite of the outcome variables. The two sets of structure *r*'s are reported in Table 4.

Table 4
Structure r's for the two LDFs.

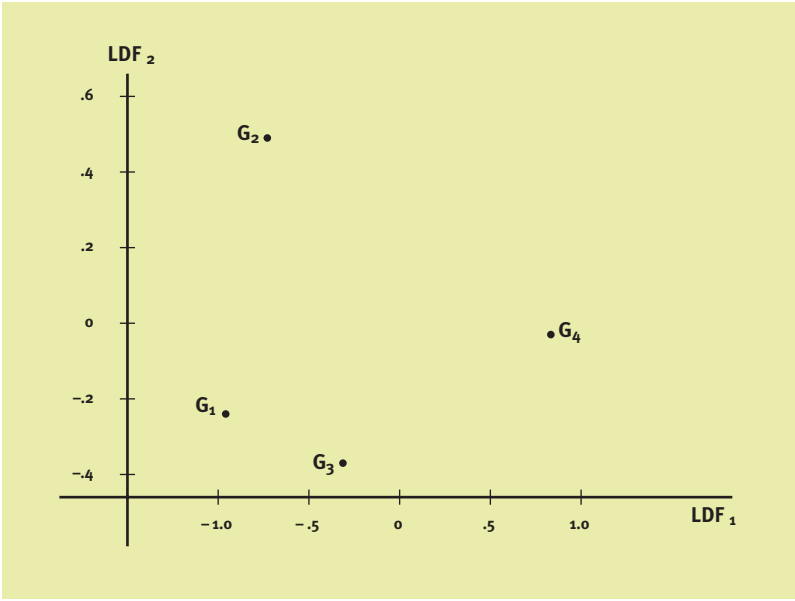
Variable	LDF1	LDF2
LINFO	**.46**	.19
SINFO	**.50**	.08
EPROF	.27	–.07
MRSNG	.23	**.71**
VTDIM	.20	**.58**
MINFO	**.68**	.23
CPSPD	.08	.20
PSINT	**.39**	.30
LLINT	.31	–.14
BMINT	.25	–.02
CMINT	.13	–.09
TRINT	–.24	.34
SOCBL	.14	–.01
IMPLS	.07	.04
MATRP	**.41**	.11

Based on these values, it may be concluded that LDF_1 is defined primarily by MINFO and secondarily by SINFO, LINFO, MATRP, and PSINT. Looking at the definitions of these five variables, one might conclude that LDF_1 represents the construct 'cognitive processes in mathematics, social science, and literature along with interest in physical science and a mature personality'. On the other hand, LDF_2 represents the construct 'cognitive processes in mathematics reasoning and visualization in three dimensions'.

LDF plot

A clearer picture of which group separation may be attributed to which LDF may be found by examining an LDF plot — see Figure 1. (The LDF group centroïds are outputted by SPSS DISCRIMINANT and SAS CANDISC). From this plot we see that LDF_1 separates Group 4 from the other three groups, while LDF_2 separates Group 2 from the other three groups. We have a substantive interpretation of the student characteristics to which we may attribute separation.

Figure 1
Plot of group centroids in LDF space.



In some multiple-group research situations it may be of interest to investigate group contrasts in addition to, or as opposed to, an investigation of the *k* overall group separation. Any contrast analysis is, in effect, a two-group analysis. For example, one may want to study all pair-wise contrasts — with the current data set, this would involve six analyses. Or, looking at Figure 1, one may want to compare Group 4 against, collectively, the other three groups; and, maybe, Group 2 against the other three. In any contrast analysis there would be one LDF. The single set of structure *r*'s would be examined to determine a 'name' for the construct to which we would attribute the resulting difference — if, of course, it was concluded by the statistical test and the effect size that there is a contrast difference. See [Huberty, 1994m] for a discussion regarding conducting contrast analyses.

It should be noted that if it is concluded that all of the covariance matrices are not 'in the same ballpark', then there is no known DDA method to use to define any meaningful LDFs. However, in this situation there may be some contrast analyses in which the subset of covariance matrices are reasonably 'close'. If so, then meaningful LDFs may be obtained.

Variable ordering

There is another bit of information in a DDA context that may be of interest. This pertains to outcome variable ordering — for which there are two approaches. One, it may be of interest to conclude which variables contribute a lot to group separation — overall or contrast-wise — and which variables contribute little to group separation. A reasonable approach, to me at least, is to conduct *p* DDAs, each with *p*-1 outcome variables. With this approach one would examine the

Wilks lambda values for each of the *p* analyses to determine a ranking (invariably with some tied ranks) of the outcome variables. The second approach to variable ordering pertains to contribution to construct definition. Such an ordering may be based on the magnitude of the structure *r*'s. For example, with regard to overall group separation, from Table 4 we see that the most important variable with respect to defining the first construct is MINFO, while MRSNG is the most important variable in defining the second construct.

Variable deletion

It may be noted that variable deletion in a DDA context is, to me at least, not too sensible. The reason that I conclude this is when designing a group separation study considerable thought and study goes in to the initial choice of outcome variables to be studied. If so, why delete some? It may very well be informative to conclude that a variable does not contribute to group separation or to construct definition. It may not be too sensible to reanalyze a subset of the original set of outcome variables.

For more details on variable deletion and variable ordering in a DDA context, see Huberty, 1994n].

Some specifics

What has been discussed in the two preceding sections does not completely cover some specific topics (and issues) related to discriminant analysis. Four topics will now be reviewed: design, alternative analyses, PDA versus DDA, and problems in reporting.

Design

For either a PDA situation or a DDA situation, it is imperative that the grouping variable(s) be well defined. This is sometimes not a seriously considered aspect of a study design. An example of poor design would be to simply trisect a continuum of a continuous variable to define three groups — for example, low, middle, and high performers on an academic assessment. Students 'near' the two cut-points need to be deleted to have a clearly defined grouping variable. In general, the basic question is: Is initial group membership clear? A second design consideration pertains to measurement of the response variables. The point here is that such measurements are expected to be reliable (to a fair extent); such reliability (usually internal consistency) should be reported.

The third design consideration pertains to sample size. Minimum group size for a PDA depends upon the analysis approach — linear or quadratic — used, the number of response variables, and upon the 'expected result'. As in many other data analysis contexts, some minimum sample size rules-of-thumb have been proposed for PDA (and for DDA). For a linear external PDA, it is recommended that the smallest group have at least 3 *p* analysis units if a 'high' hit rate of inter-

est is expected. If a 'low' hit rate is expected (based on previous research and/or substantive knowledge), the minimum may be 5 *p*. When a quadratic external rule is used, the recommended desired minimums are 5 *p* and 7 *p* for high and low expected hit rate, respectively. For a DDA, the desired minimum sample size depends upon the number of outcome variables (i.e. *p*), and upon a guesstimate of the effect size. One may relate the DDA sample size issue to MANOVA. One rule of thumb is for the minimum group size to be 4 *p* or 5 *p* for a 'moderate' effect size [Huberty, 1994o]. (An effect size in MANOVA may be expressed as unadjusted eta-squared value.)

Alternative analyses

The PDA approach discussed earlier in this chapter has been termed the 'normal-based' approach. This term is used because the posterior probability of group membership is based on multivariate normality [Huberty, 1994p]. There are, however, some alternative approaches to predicting group membership that have received some support. Some alternatives are logistic classification, probit classification, nearest-neighbor classification, neural networks, recursive partitioning, and others. Excellent sources for these alternatives (and more) are [Hand, 1997; Krzanowski, 1995; McLachlan, 1992; Rencher, 1995; Rencher, 1998]. A rather mundane discussion of non-normal classification rules is given by [Huberty, 1994q], including the handling of categorical predictor variables, and the handling of predictor scores that are ranked.

As mentioned earlier in this chapter, internal classification — where one data set is used to determine the rule as well as to determine hit rate estimates — is not the recommended approach to estimate hit rates (unless, of course, the *n*/*p* ratio is extremely large). Rather, an external rule is recommended. What was herein suggested is the leave-one-out (L-O-O) method. This is a hybrid of the well-known re-sampling method, the jackknife. Another type of re-sampling that has been applied in hit rate estimation is the bootstrap [Hand, 1997a]. With regard to the L-O-O method of hit rate estimation, [Hand, 1997b] concludes that "the relatively large variance of this method has now driven it out of favor and bootstrap methods seem to be preferred....".

It has become increasingly popular, at least in the behavioral sciences, to report a value of what has come to be termed an effect size index. The interpretation of such an index varies somewhat across research disciplines. Such an index value is common for designs with two or more groups of analysis units. Invariably, the univariate context in which effect sizes are discussed is when the equal-variance condition is met. The unequal-variance condition is virtually ignored. Here is where a general effect-size approach needs to be considered. What is suggested by [Huberty, 2000] is to use the *I* index (discussed earlier in this chapter) as an index of effect size. It is noteworthy to mention that the *I* index may be

		PDA	DDA
1	Research concern	Prediction of group membership	Description of group separation
2	Variable roles:		
	Predictor(s)	Response variables	Grouping variable
	Criterion (ia)	Grouping variable	Response variables
3	Response variable set	Hodgepodge	System
4	Response variable composite	LCF	LDF
5	Number of composites	k	min (p, k-1)
6	Preliminary analysis concerns:		
	Equality of covariance matrices	Yes	Yes
	MANOVA	No	Yes
7	Analysis aspects of typical interest:		
	Variable construct(s)	No	Yes (!)
	Response variable deletion	Yes (!)	Maybe
	Response variable ordering	Yes	Yes
8	Criterion for variable deletion/ordering	Classification accuracy	Group separation
9	Research purpose	Practical/theoretical	Theoretical
10	Interest in generalizability	Yes	Yes

Table 5
PDA versus DDA. Context: k groups of units, p response variables.

used under the condition of equal or unequal variances — or unequal covariance matrices in the multivariate context.

PDA versus DDA

It is felt by the current writer, at least, that the distinction between a research situation calling for a PDA and a research situation calling for a DDA should be recognized by a researcher who uses either. The basic distinction has been alluded to earlier in this chapter. A summary of distinguishing aspects of the two analyses is presented in Table 5.

Problems in reporting

The reporting of results of a PDA and of a DDA has not been very complete. And in some cases reporting has been quite misleading! A detailed discussion of problems in reporting results of discriminant analyses is given by [Huberty and Hussein, 2000].

References

- Fisher, R.A. (1936). The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics **7**:179-188
- Hand, D.J. (1997a). Construction and Assessment of Classification Rules. Wiley & Sons, New York. pp123-125

– Hand, D.J. (1997b). Construction and Assessment of Classification Rules. Wiley & Sons, New York. p126
– Huberty, C.J. (1994). Applied Discriminant Analysis. Wiley & Sons, New York. –a: p278, -b: p64; -c: p59; -d: p368; -e: pp102-108; -f: p107; -g: pp76-77, pp79-80; -h: Chapter VIII; -i: pp217-222; -j: pp182-189; -k: Chapter XV; -l: p213; -m: pp196-200; -n: Chapter XVI; -o: p201; -p: pp56-57; -q: Chapter X
– Huberty, C.J., L.L. Lowman. (1998). Discriminant Analysis in Higher Education Research. In: J.C. Smart. (ed.). Higher Education: Handbook of Theory and Research. Agathon Press, New York. pp181-234
– Huberty, C.J., M.H. Hussein. (2000). Some Problems Reporting Use of Discriminant Analyses. 22nd Biennial Conference of the Society for Multivariate Analysis in the Behavioural Sciences, London
– Huberty, C.J., L.L. Lowman. (2000). Group Overlap as a Basis for Effect Size. Educational and Psychological Measurement **60**:543-563
– Huberty, C.J., M.D. Petoskey. (2000). Multivariate Analysis of Variance and Covariance. In: H.E.A. Tinsley, S.D. Brown. (eds.). Handbook of Applied Multivariate Statistics and Mathematical Modeling. Academic Press, New York. pp183-208
– Krzanowski, W.J., F.H.C. Marriott. (1995). Multivariate Analysis. Part 2. Classification, Covariance Structures and Repeated Measurements. Arnold, London
– McLachlan, G.J. (1992). Discriminant Analysis and Statistical Pattern Recognition. Wiley & Sons, New York
– Morrison, D.F. (1990). Multivariate Statistical Methods. Chapter 6, McGraw-Hill, New York
– Rencher, A.C. (1995). Methods of Multivariate Analysis. Wiley & Sons, New York
– Rencher, A.C. (1998). Multivariate Statistical Inference and Applications. Wiley & Sons, New York

6.2.4 Subspace methods

Dick de Ridder[1]

Introduction

Often, an object or a process can be characterized by a large number of measurements. When (statistical) algorithms are applied, each measurement corresponds to a dimension in a measurement space. For example, in an application separating the good from the bad apples, measurements may be color, width, height, surface properties, etc.; each of which defines one or more dimensions. However, for many applications far fewer dimensions suffice to describe the data accurately. Moreover, statistical algorithms require a large set of examples to estimate parameters in these high-dimensional spaces. For example, when image data is used directly, each individual pixel can be considered a dimension; but it is impractical and unnecessary to describe it in this way, as even small images lead to tens of thousands of dimensions.

What is needed is a way of mapping the data to a smaller number of dimensions without losing (too much) information. Methods for doing so have been studied extensively in, amongst others, pattern recognition literature:

- feature selection, which discards individual measurements;
- multidimensional scaling, which embeds the original data's distance matrix in a lower-dimensional space;
- subspace methods, which combine the original measurements to a smaller number of dimensions. Such subspaces can be linear or non-linear. Examples of methods to find non-linear subspaces are principal curves, projection pursuit and diabolo neural networks; however, non-linear subspaces are beyond the scope of this section.

Linear subspaces, the simpler case, have a well-defined mathematical form: they can be completely characterized by their origin and basis vectors. As an example, Figure 1 shows a 2D subspace in a 3D space. Discussed below are two different methods of finding linear subspaces.

- Principal Component Analysis (PCA), sometimes called the Karhunen-Loève transformation (KL), is a widely used subspace method. The subspace maximizes the amount of variance retained and de-correlates the data.
- Independent Component Analysis (ICA) is a relatively new method of reducing the number of dimensions which demands that the distributions of projected data are independent.

Besides lowering the computational load of further data processing and lowering the number of examples needed for statistical algorithms, subspace meth-

1 Dr Ir D. de Ridder, dick@ph.tn.tudelft.nl, Pattern Recognition Group, Faculty of Applied Sciences, Delft University of Technology, Delft, The Netherlands, http://www.ph.tn.tudelft.nl/

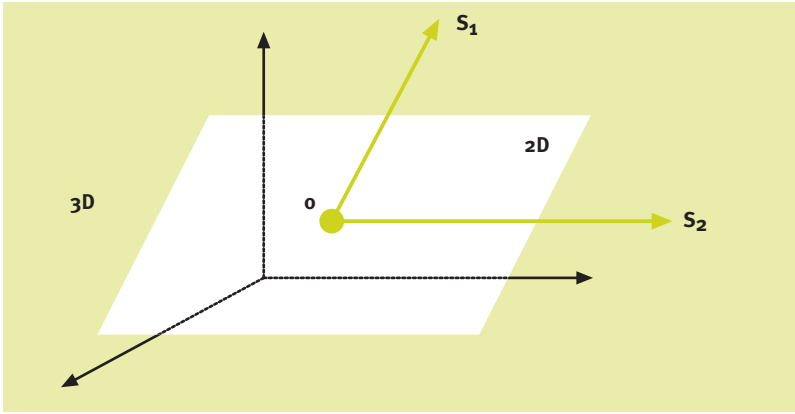


Figure 1
A 2D linear subspace in 3D, defined by its origin O and two basis vectors S_1 and S_2.

ods can also be extremely useful for explanation (which measurements are useful?) and data visualization.

Principal Component Analysis

Principal Component Analysis (PCA) is already quite old, proposed in 1901 by Pearson; in 1935, Hotelling gave a practical computing method. This section will not go into too much detail on PCA, as there is a large body of literature on it (see the standard textbooks on pattern recognition in the reference section). Some attention will be paid to a relatively new development, probabilistic PCA, which gives PCA a full probability model and allows training of mixtures – of PCAs.

Model

PCA is a linear projection technique:

$$u = W(x - \mu)$$

where μ is the *m*-dimensional projected data vector, x is the original *d*-dimensional data vector and the matrix W contains the PCA projection vectors. Usually, $m << d$. PCA demands that projection vectors stored in W:

- maximize the variance retained in the projected data;
- or (equivalently) give uncorrelated projected distributions;
- or (equivalently) minimize the least square reconstruction error.

As an example, Figure 2(a) shows the first (longest) and second basis vector found by PCA on a 2D data set.

PCA is closely related to factor analysis (FA). The difference is that PCA assumes noise (outside the subspace) to have identical variance, whereas FA assumes

different dimensions have different noise variance. Contrary to PCA, there is no closed-form solution for FA.

PCA algorithm

It is not very difficult to show that the m projection vectors maximizing the variance of u, i.e. the principal *axes*, are given by the eigenvectors $\mathbf{e}_1 \ldots \mathbf{e}_m$ of the data set's covariance matrix $\mathbf{C}$, corresponding to the m largest non-zero eigenvalues $\lambda_1 \ldots \lambda_m$.

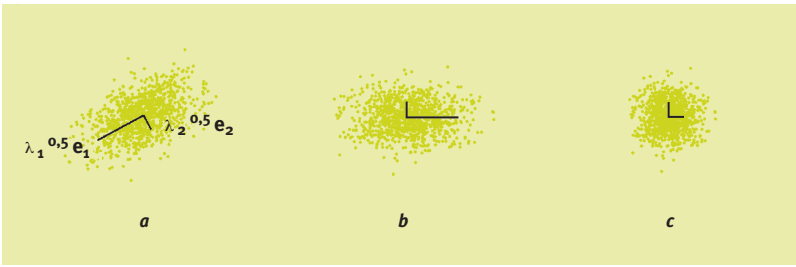


Figure 2
a *PCA basis vectors;*
b *after PCA projection;*
c *after sphering.*

The data set's covariance matrix can be estimated as:

$$\hat{C} = S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T$$

and the eigenvectors and -values can be found by solving the set of equations:

$$(S - \lambda_i I)e_i = 0, \qquad \forall i = 1,\ldots,d$$

After calculating the eigenvectors, sort them by their corresponding eigenvalues and select the m vectors with the largest eigenvalues. The PCA projection matrix is then $W = E^T$, where E has the selected eigenvectors as its columns.

Dimensionality reduction with PCA

The question is: will m have to be set by hand, or is there some heuristic way of finding a good value for it? Well, the proportion of variance retained by mapping down to m dimensions can be found as the normalized sum of the m largest eigenvalues:

$$r = \left(\sum_{i=1}^{m}\lambda_i \Big/ \sum_{i=1}^{d}\lambda_i\right) \times 100\%$$

In many applications, m is chosen such that at least $r = 90\%$ or $r = 95\%$ variance is retained. The remaining variance is assumed to be due to noise.
As an example, consider a data set containing 1,000 16 x 16 pixel images each of 10 digits, stored as vectors. Performing PCA on the covariance matrix of the 10,000 256-dimensional vectors gives Figure 3. Preserving 90% of the variance

leaves only 50 dimensions; but one could also choose to use just 25 dimensions, leaving roughly 80% of the variance. 50-100 is called the intrinsic dimensionality of the data: it may look as if it has 256 dimensions, but it really only 'lives' in a 50D-100D subspace.

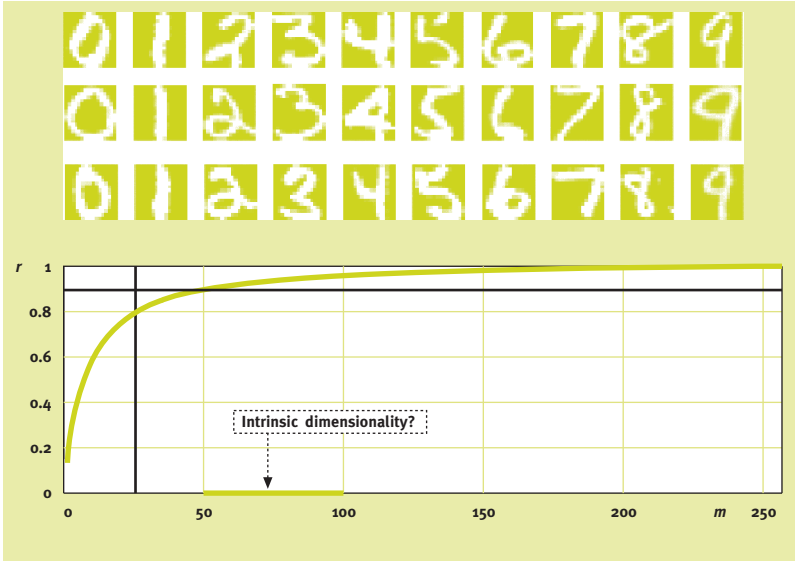


Figure 3
Intrinsic dimensionality.

Other properties of PCA

Besides retaining the maximum amount of variance in the projected data, PCA also has the following properties:

– *de-correlation*: the projected data u is de-correlated, i.e. the covariance matrix $E(uu^T)$ is a diagonal matrix. Figure 2(b) illustrates this; given the data set above, it shows the data projected onto both eigenvectors. The variance in each principal direction can also be normalized, by using

$$W = \Lambda^{-\frac{1}{2}} E^T$$

instead of just $W = E^T$ (where Λ is a matrix containing the eigenvalues on the diagonal). In this case, the covariance matrix of u is equal to the identity matrix I. Using PCA in this way is called sphering or whitening, illustrated in Figure 2(c).

– *least squares reconstruction*: if the projected vectors u are projected back into the original space using

$$\hat{x} = AW(x - \mu)$$

with $A = W(WW^T)^{-1}W^T$, then the reconstruction error:

$$\left\| (x - \mu) - \hat{x} \right\|$$

is minimal.

Probabilistic PCA

Probabilistic PCA (PPCA) is an extension of traditional PCA, proposed by Roweis, and Tipping and Bishop. It defines a proper probability model for PCA. Note that in traditional PCA, directions 'outside' the subspace are simply discarded. In PPCA however, these directions are assumed to contain Gaussian noise[2]. Furthermore, the original data x is modeled as being generated by lower-dimensional data u:

$$x = Au + \mu + \varepsilon$$

where $\boldsymbol{\varepsilon}$ is Gaussian noise with covariance matrix $\sigma^2 I$. This is called a latent variable model: the latent variables u 'cause' the observed variables x. The entire data set is then modeled by a Gaussian with restricted covariance matrix:

$$p(x) = \frac{1}{(2\pi)^{d/2} |\det C|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)\right)$$

where $C = W^T W + \sigma^2 I$.

W is found as in the original PCA algorithm, and σ^2 is found by calculating the average of the variance in the discarded directions:

$$\sigma^2 = \frac{1}{d-m} \sum_{i=m+1}^{d} \lambda_i$$

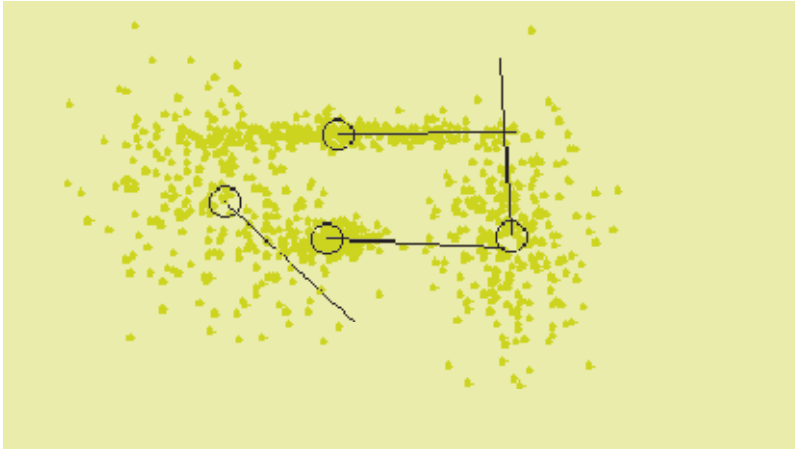
The advantage of PPCA over traditional PCA is that it defines a proper probabilistic model. This model can easily be extended to mixture models (see below). Bishop has also proposed Bayesian methods to automatically determine *m*, the number of dimensions to retain. A disadvantage of PPCA is that it needs to estimate an extra parameter, σ^2. In cases where there's little noise outside the subspace, σ^2 will be extremely small and training is very difficult or impossible.

Mixtures of PCA

A limitation of PCA is that it can only model linear subspaces. Often, however, the data lies in a non-linear subspace. In this case, it can be useful to model the data using multiple PCA subspaces, each of which is responsible for just part of the data. Figure 4 shows an example. To fit a mixture of ordinary PCAs, general iterative clustering algorithms can be used to divide the data among the sub-

2 Noise with a Gaussian distribution, a distribution with a zero mean and a variance σ.

Figure 4
A mixture of 4 1D PCA subspaces.



spaces; for probabilistic PCA, Tipping and Bishop formulated an expectation-maximization (EM) algorithm.

Applications of PCA

PCA can be applied in almost any field in which large datasets play a role: multivariate statistics in general, psychology, biology, medicine, the social sciences, etc. In most of these applications it is used to allow visualization of data with more than 3 dimensions; to remove noise from data; or to find which combinations of measurements describe the data optimally (i.e. to model the data).

In pattern recognition, PCA is often used to preprocess data in order to remove unnecessary dimensions and noise. Also, PCA can be a reasonable model for high-dimensional data. For example, when images of faces are treated as points in very high-dimensional spaces spanned by all pixels (e.g. a 256 x 256 image is a point in a 65,536D space), PCA can be used to remove all but, say, 100 dimensions. The resulting basis vectors can again be visualized as images and are called eigenfaces.

Independent Component Analysis

Independent Component Analysis (ICA) is a method related to PCA and projection pursuit, which finds a set of basis vectors for a (sub)space based on a data set. ICA demands that projections of the data projected onto each of the basis vectors are independent. Independent components are often found by looking for directions in which the data has a non-Gaussian distribution (because of the central limit theorem, sums of random variables have a Gaussian distribution, so non-Gaussian distributions are likely to correspond to single distributions, not sums).

Below, the ICA model will be discussed and compared to PCA. Algorithms to find

ICA bases will be introduced, and some applications will be given. For more information, try one of Hyvärinen's survey papers (see Reference Section).

Model

ICA is a linear projection technique, just like PCA. The basic model assumes that the data, x, is a linear combination of a number of unknown sources s:

$$x = As + \mu$$

Here A is the unknown *mixing matrix* and μ is the origin of the data. The inverse model is:

$$u = W(x - \mu)$$

where u is the estimate of s and W is the inverse of A, the *unmixing matrix*. Most algorithms deal with the case where the number m of basis vectors sought is equal to the number of dimensions d in the data, so $W = A^{-1}$. However, there are also algorithms for finding undercomplete bases ($m < d$) and overcomplete bases ($m > d$).

ICA versus PCA

The goal of ICA is — given only x — to find both W and u. There are infinitely many solutions, so an added demand on the source estimates u is that they are independent. This goes further than PCA, which just demands that they are uncorrelated.

Figure 5 illustrates the difference between ICA and PCA: the data consists of 2D uniformly distributed samples. Both methods find a new set of basis vectors for the data. But, where PCA maximizes the variance and projections onto the basis vectors are mixtures, ICA correctly finds the two vectors onto which the projections are independent.

ICA algorithms

Contrary to PCA, there is no closed form expression to find an ICA base. Instead, many iterative algorithms have been proposed based on the following observations:

– independent source distributions should have a smaller differential entropy than the Gaussian distribution;

$$H(u) = -\int f(u) \ln f(u) \, du$$

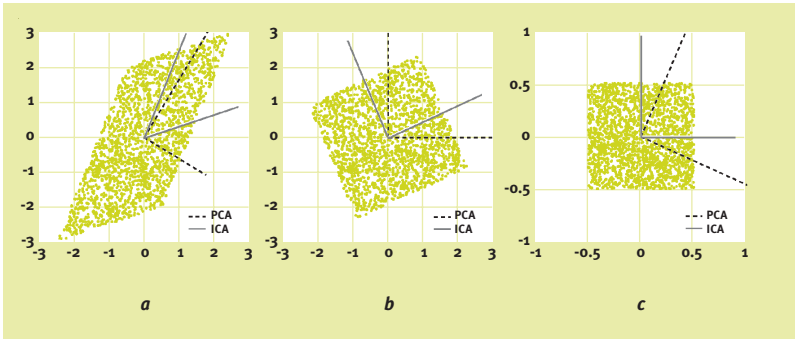


Figure 5
a *PCA and ICA basis vectors;*
b *after PCA projection;*
c *after ICA projection.*

independent sources u_i should have as little mutual information, i.e. about each other as possible;

$$I(u_1,\ldots,u_m) = \sum_{i=1}^{m} H(u_i) - H(\mathrm{u})$$

– the Kullback-Leibler divergence between the factorized density and the true density $f(\mathrm{u})$ should be minimal;

$$f'(\mathrm{u}) = \prod_{i=1}^{m} f_i(u_i)$$

$$\int f'(\mathrm{u}) \ln \frac{f'(\mathrm{u})}{f(\mathrm{u})}\, d\mathrm{u}$$

– independent sources are likely to be found by looking for non-Gaussian distributions in the projection, e.g. by specifying a non-Gaussian distribution and fitting it using maximum likelihood.

Most of the observations listed above lead to similar or even identical algorithms. The general idea behind all of them is that distributions of the data projected onto an ICA basis vector should be as non-Gaussian as possible. This links ICA to projection pursuit, which often uses non-Gaussianity as a measure of 'interestingness' of a projection. An intuitive reasoning is that, due to the central limit theorem, which states that sums of random variables will tend in the limit to have a Gaussian distribution, non-Gaussian projection distributions will indicate that the projection is not a sum of random variables, but a single one.

Measuring non-Gaussianity

A measure often used to judge the property of non-Gaussianity is the (Pearson) kurtosis of a distribution $f(\mathrm{u})$, i.e. the central fourth-order moment:

$$\kappa_{f(u)} = \frac{\mu_4}{\mu_2^2} = \frac{\mathrm{E}\left(u'^4\right)}{\mathrm{E}\left(u'^2\right)^2}$$

where $u' = u - E(u)$. Figure 6 shows the kurtoses from data drawn from three types of distribution.

A limitation of ICA therefore is that Gaussian independent components cannot be found. However, for Gaussian ICs simple de-correlation (as performed by PCA) also makes the components independent, as the Gaussian distribution is specified completely by the mean and covariance matrix. When it is known that data is distributed according to a Gaussian, whitening will result in Gaussian projections.

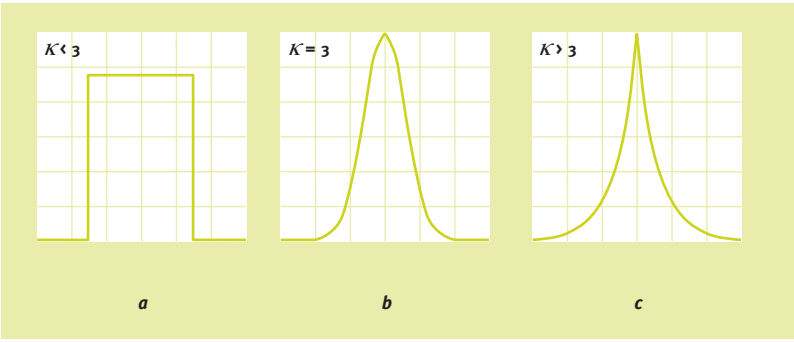


Figure 6
a *uniform distribution;*
b *Gaussian distribution;*
c *Laplacian distribution.*

Applications of ICA

The first major application of ICA was to the blind unmixing of signals (also known as blind source separation or the cocktail party problem). The idea is that when measurements are made using sensors placed at various locations (for example in noise measurements around airports, or in EEG measurements in medicine), each sensor picks up a different combination of a number of sources. ICA can then be used to identify the individual sources. A simple demonstration is shown in Figure 7. In image processing, ICA can be applied to compression, de-noising and separating spectra in hyperspectral images. ICA also gives interesting decompositions of images shown to be related to visual receptive fields in mammals, indicating a possible link with biological principles. Figure 8 illustrates this.

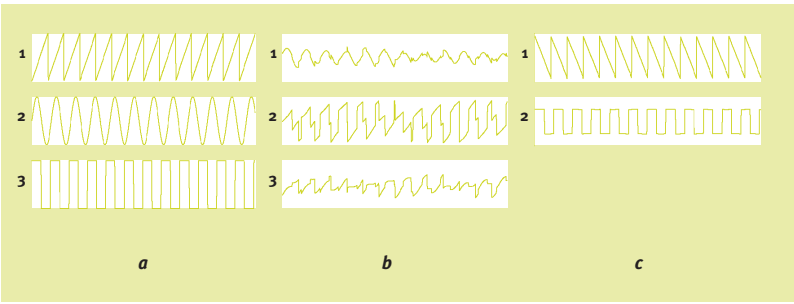


Figure 7
a *3 original signals;*
b *3 mixed signals;*
c *2 ICA unmixed signals.*

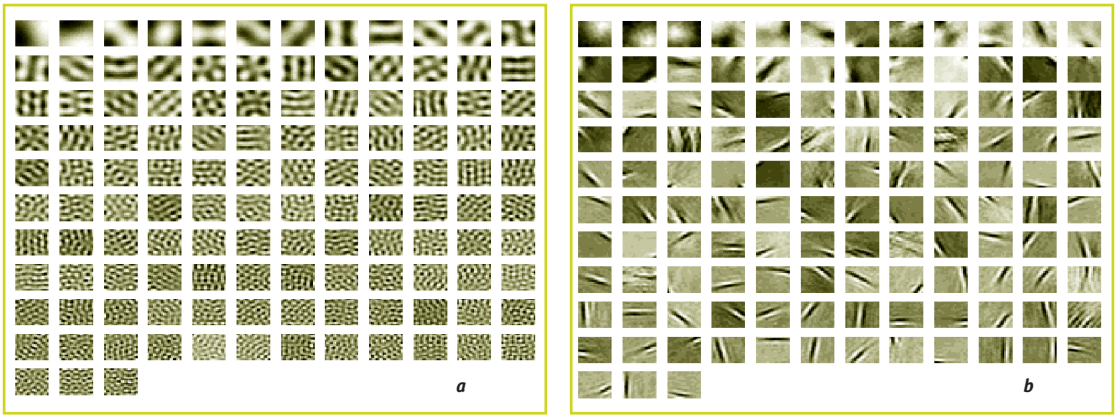


Figure 8
a *12x12 pixel PCA basis vectors found from natural images;*
b *ICA basis vectors on same images.*

Summary

PCA and ICA are two methods of describing data in subspaces, using fewer dimensions than the original number of measurements. Although both have the same model, the different demands on the projected data lead to different algorithms. PCA subspaces can be calculated quite quickly, whereas iterative algorithms have to be used to find ICA subspaces. Both can be applied to lower the number of dimensions of data sets in order to visualize or preprocess for further statistical algorithms. They can also be used to model the data. In this case, PCA has more widespread applicability than ICA, which is mainly useful for certain specific problems such as blind source separation.

References

- Bell, A.J., T.J. Sejnowski. (1995). An Information-Maximization Approach to Blind Separation and Blind Deconvolution. Neural Computation **7** (6):1129-1159. http://sloan.salk.edu/~tony/ica.html
- Bell, A.J., T.J. Sejnowski. (1997). Edges are the 'Independent Components' of Natural Scenes. In: M.C. Mozer, M.I. Jordan, T. Petsche. (eds.). Advances in Neural Information Processing Systems **9**:831. MIT Press, Cambridge, MA. http://sloan.salk.edu/~tony/ica.html
- Bishop, C.M. (1999). Bayesian PCA. In: S.A. Solla, T.K. Leen, K.-R. Müller. (eds.). Advances in Neural Information Processing Systems **12**:382-388. MIT Press, Cambridge, MA
- Devijver, P.A., J. Kittler. (1982). Pattern Recognition, a Statistical Approach. Prentice Hall, London
- Field, D.J. (1987). Relations between the Statistics of Natural Images and the Response Properties of Cortical Cells. Journal of the Optical Society of America A **4** (12):2370-2393
- Friedman, J.H. (1987). Exploratory Projection Pursuit. Journal of the American Statistical Association **82** (397):249-266

- Hyvärinen, A. (1999). Sparse Code Shrinkage: Denoising of NonGaussian Data by Maximum Likelihood Estimation. Neural Computation **11** (7):1739-1768. http://www.cis.hut.fi/~aapo
- Hyvärinen, A. (1999). Survey on Independent Component Analysis. Neural Computing Surveys **1** (2):94-128. http://www.cis.hut.fi/~aapo
- Hyvärinen, A., E. Oja. (2000). Independent Component Analysis: Algorithms and Applications. Neural Networks **13** (4-5):411-430. http://www.cis.hut.fi/~aapo
- Kaarna, A., P. Zemcik, H. Kälviäinen, J. Parkkinen. (1998). Multispectral Image Compression. Technical Report Research Report No. 60. Department of Information Technology, Lappeenranta University of Technology, Lappeenranta, Finland. http://www.it.lut.fi/research/ip/projects/mic/mic_rep.ps.gz
- Lee, T.-W., M. Girolami, T.J. Sejnowski. (1999). Independent Component Analysis Using an Extended Infomax Algorithm for Mixed Sub-Gaussian and Super-Gaussian Sources. Neural Computation **11** (2):417-441. http://www.cnl.salk.edu/~tewon/pubs.html
- Lewicki, M.S., T.J. Sejnowski. (2000). Learning Overcomplete Representations. Neural Computation **12** (2):337-365. http://www.cs.cmu.edu/~lewicki
- Manly, B.F.J. (1994). Multivariate Statistical Methods, a Primer. 2nd Edition. Chapman & Hall, London
- Ridder, D. de. (2001). Adaptive Methods of Image Processing. PhD Thesis. Delft University of Technology, Delft. http://www.ph.tn.tudelft.nl/~dick
- Roweis, S. (1997). EM Algorithms for PCA and SPCA. In: M.I. Jordan, M.J. Kearns, S.A. Solla. (eds.). Advances in Neural Information Processing Systems **10**. MIT Press, Cambridge, MA. http://www.gatsby.ucl.ac.uk/~roweis/publications.html
- Tipping, M.E., C.M. Bishop. (1997). Probabilistic Principal Component Analysis. Technical Report Technical Report NCRG/97/010. Neural Computing Research Group, Aston University, Birmingham, UK. http://www.gatsby.ucl.ac.uk/~quaid/course/readings/ppca.ps
- Tipping, M.E., C.M. Bishop. (1999). Mixtures of Probabilistic Principal Component Analyzers. Neural Computation **11** (2):443-482. ftp://ftp.research.microsoft.com/users/mtipping/mppca.ps.gz
- Turk, M., A.P. Pentland. (1991). Eigenfaces for Recognition. Journal of Cognitive Neuroscience **3** (1):71-96

6.2.5 Introduction to multidimensional scaling

Elżbieta Pękalska[1]

Multidimensional scaling (MDS) refers to a group of methods in the area of multivariate data analysis. Originally, the theory of MDS was developed in the behavioral and social sciences for studying the relations between objects. Since then, MDS has become more popular and its applications have been extended to other disciplines, like accounting, marketing, sociology, education, psychology or pattern recognition.

If the data consists of one or two variables, then many simple methods are available for its visualization, showing or emphasizing some properties or relations between objects. When multivariate data with e.g. six, ten or more variables (features) has to be examined, then the human limitation to operate in at most 3 dimensions makes it hard for them to judge whether the data contains any structure. So, there is a need for smarter, more sophisticated techniques.

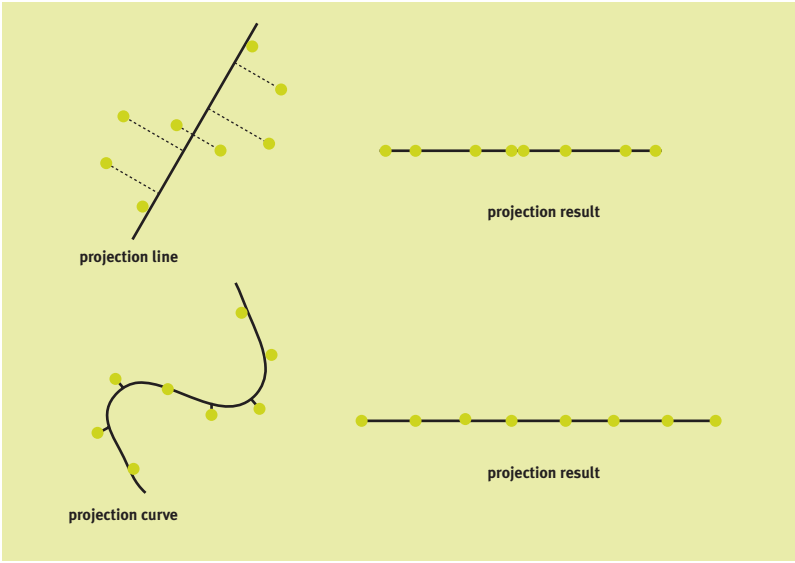


Figure 1
The difference between linear and non-linear projection methods.

An early stage of the exploration of a multivariate dataset is to visualize it on a plane or in a 3-dimensional space. By doing this, some intuition about the data can be gained, as well as understanding of the relations between objects, the intrinsic structure or possible cluster tendencies, etc. In fact, the goal is to represent the data in a low, 2- or 3-dimensional space so that the total configuration could reflect important relations between the individual points. MDS techniques make this possible.

Methods facilitating the visualization of multivariate data in a low-dimensional space are called projection methods. They can also be considered as methods

1 Dr E. Pękalska, ela@ph.tn.tudelft.nl, Pattern Recognition Group, Faculty of Applied Sciences, Delft University of Technology, Delft, The Netherlands

for dimension reduction, not necessarily to a few dimensions. There are linear and non-linear projection methods. The main difference between them is that a linear method looks for a linear subspace (like a line or a plane) to project the data onto, while a non-linear technique searches for a non-linear subspace. The latter can be envisioned as wrapping an elastic surface around points in a high-dimensional space and then projecting the points onto it (see Figure 1). In this way, much more flexibility can be introduced for adding extra conditions or constraints to be fulfilled in a projection process, for instance, the preservation of all (or some particular) distances. In Figure 1, the difference between linear and non-linear projection is illustrated by a 2-dimensional example. The points are spread equidistantly, but after projection onto a line by a linear method the distances between points are only roughly preserved. In case of a non-linear technique, a non-linear 1-dimensional curve can be found such that the nearest neighbor distances are preserved.

MDS is carried out on data relating to subjects or objects, which are characterized by (dis)similarity measurements describing the relationships between them. A similarity value indicates the degree of resemblance between two objects. A dissimilarity value describes how much the objects differ. The larger the dissimilarity value, the less similar the objects are. For simplicity, however, we will further refer to dissimilarities only, since similarities can easily be transformed into dissimilarities [Borg, 1997; Cox, 1995].

The purpose of MDS is to obtain a representation of dissimilarity data in a low-dimensional space so that data can be accessible for visual inspection and exploration or as an analytic approach, in order to discover rules that will help to explain the data. The output of MDS is a spatial representation of the data, consisting of a configuration of points, representing the objects, in a low-dimensional space. Such a display is visually appealing to the human eye and often allows for a better comprehension of the data. The configuration is believed to reflect significant characteristics, as well as 'hidden structures' of the data. Therefore, objects judged to be similar to one another result in points being close to each other in a low-dimensional space. The larger the dissimilarity between two objects, the further apart they should be in the resulting map of points.

The MDS algorithms designed for analyzing dissimilarity data can be roughly divided into two basic types: metric and non-metric. Fully metric analysis assumes that both the input data and the output configuration are metric. Fully non-metric analysis assumes that the input data are nominal or ordinal (non-metric).

Illustrative examples

A simple example which gives an idea of what can be achieved by MDS is the reconstruction of a map of a country, given the road/air distances between main cities. Here, the road distances between 12 major cities in The Netherlands are

	R	A	U	DHg	M	E	B	A	Z	L	G	DHr
Rotterdam	–	73	57	21	202	113	51	118	147	206	251	144
Amsterdam		–	37	55	213	121	106	99	113	139	203	82
Utrecht			–	62	180	88	73	64	91	181	195	120
Den Haag				–	223	134	72	118	152	188	252	126
Maastricht					–	86	146	167	244	334	348	296
Eindhoven						–	57	82	150	240	254	204
Breda							–	111	156	246	260	189
Arnhem								–	68	158	172	182
Zwolle									–	91	105	157
Leeuwarden										–	62	90
Groningen											–	154
Den Helder												–

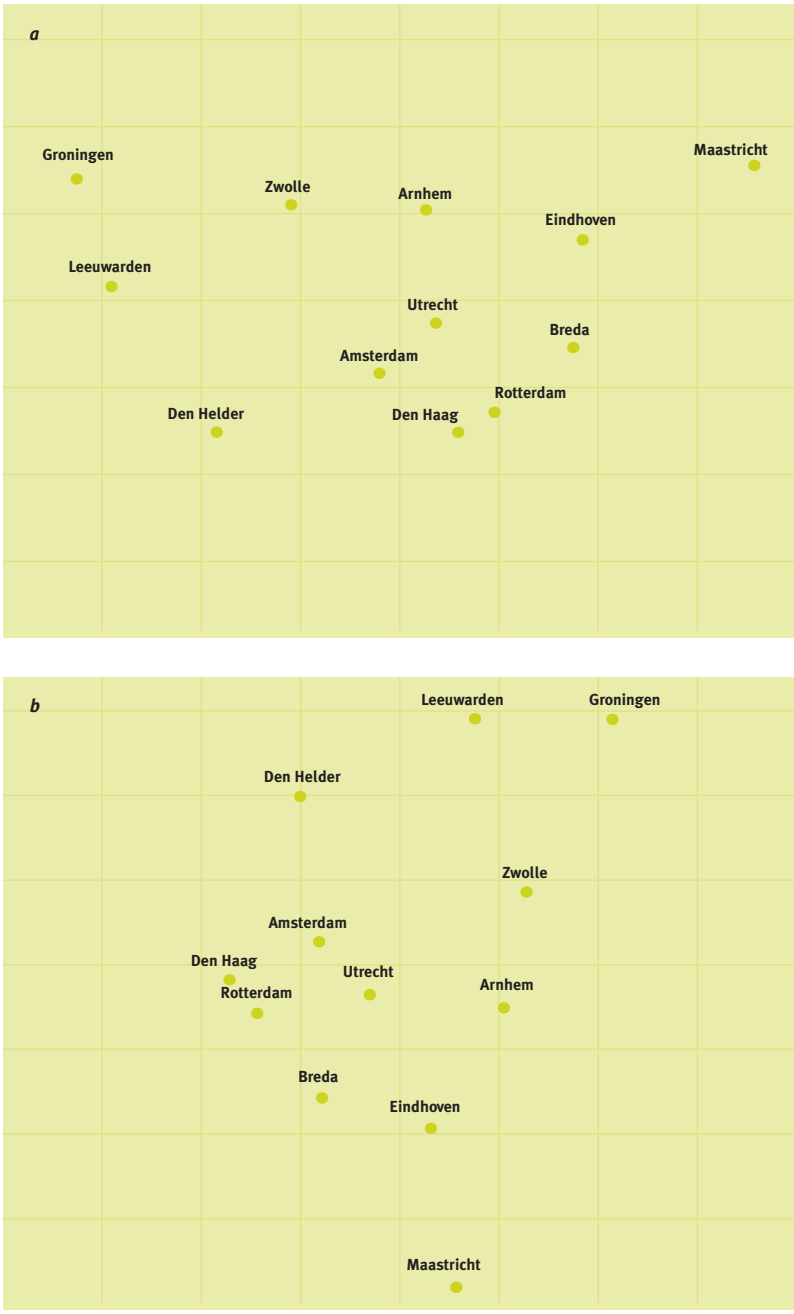
Table 1
Road distances between 12 Dutch cities. Horizontally, only the key letters of the cities' names are used.

considered, as provided in Table 1. The distances are only given for the upper part of a 12 × 12 distance matrix D, since they are symmetric. The results of the MDS mapping can be observed in Figure 2. A comparison between the original map and the result given by the MDS technique makes it clear that the MDS method has been successful in recovering the location of the cities. In general, the cities are shown in good relation to each other, with the possible exception of Den Haag and Rotterdam in relation to Amsterdam. It can be also observed in Figure 3, i.e. on the plot of the MDS distances against the original road distances. For perfectly preserved distances, we would observe points lying on a line $y=x$. In our case, they only slightly deviate from this ideal solution, which confirms our visual judgments. In general, such plots of the MDS distances versus original dissimilarities should accompany our data analysis.

There are two important things to realize about an MDS map. The first is that the axes are, in themselves, meaningless and the second is that the orientation of the picture is arbitrary. Thus, an MDS representation of distances between the Dutch towns need not be oriented so that north is up and east is right (as observed in Figure 2(a)). What is important in an MDS map, is the relative positions of the objects.

In general, MDS serves for exploration of the data, e.g. finding possible clusters, i.e. groups of points which are close together in the represented space. As an example, let us consider auditory confusion (dissimilarities) between 25 letters (all excluding 'O') and 10 Arabic numerals [Lee, 2001]. The spatial MDS map is shown in Figure 4. The cluster of similarly sounded letters or numerals can be clearly observed. For instance, we can justify the 'closeness' of 'I', '5', '1' or 'Y' since, when spoken, they sound similar to each other.

Figure 2
The reconstructed map of The Netherlands
a *original;*
b *rotated.*



Another purpose of MDS is to find rules that would explain observed dissimilarities and would help to describe the data structure in simple terms. This might be especially useful for data describing human judgments of similarity between objects. In such a case, interpreting an MDS configuration entails making a link between geometrical properties of such a map and prior knowledge about the objects represented as points [Borg, 1997]. By identifying points which are far

Figure 3
The distances of the MDS configuration of Dutch cities against original road distances.

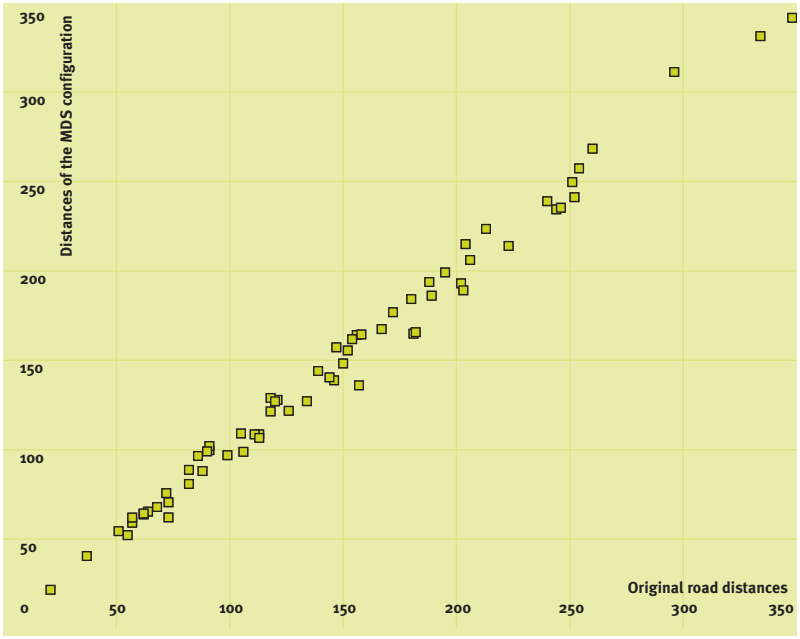
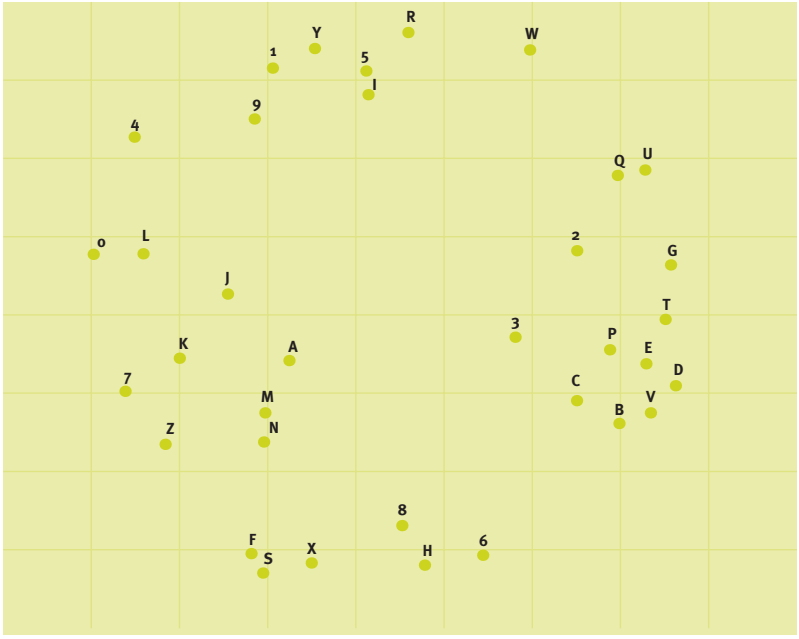


Figure 4
The spatial relations of auditory confusion measurements between letters and numerals.



apart, a line between them can be drawn, defining a perceptual axis, which describes a direction of a change between opposite or significantly different characteristics. This involves some data-guided speculation. An example is based on human judgments of similarities [Lee, 2001] between 21 sports for which the MDS representation is given in Figure 5. The dashed lines in this figure were added by us as a possible (not unique) interpretation, to aid the

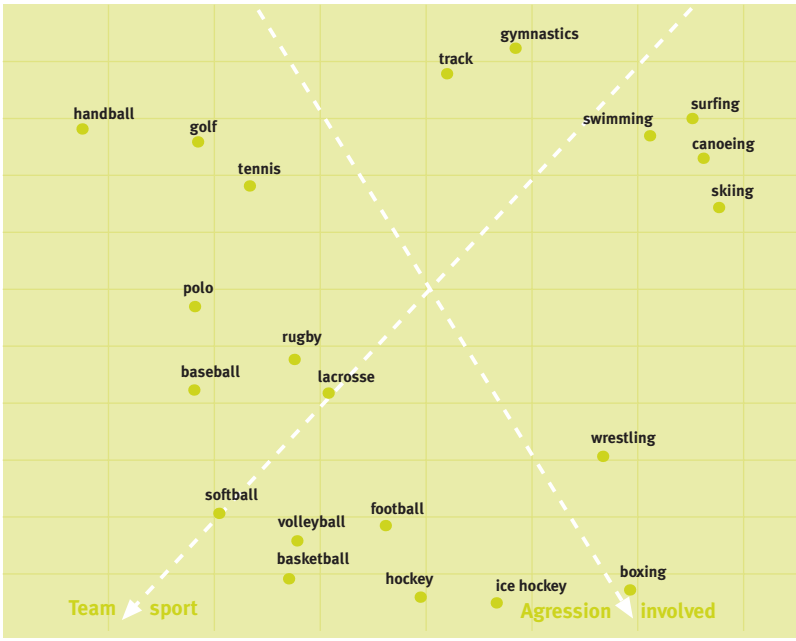


Figure 5
The MDS map of sports based on the similarity judgments by humans.

understanding of the distribution of points. To interpret why humans consider some sports to be more alike than others, we could distinguish two perceptual axes: team sport versus individual sport, and the degree of aggression involved. (Note that such axes do not need to be perpendicular at all.) Another possibility could be to evaluate the difference on the basis of whether a ball is used in a sport or not.

Dissimilarities

We assume here that the dissimilarity between an object and itself is zero, standing for perfect similarity. Dissimilarity measures, commonly used in practice, are the L_p distances, which for m-dimensional real-valued vectors are defined as:

$$d_p(x,y) = \left(\sum_{k=1}^{m} |x_k - y_k|^p \right)^{1/p}$$

where x_k and y_k are the k-th coordinates of $\boldsymbol{x}$ and $\boldsymbol{y}$. A Euclidean distance is obtained for $p=2$, i.e.

$$d(x,y) = \left(\sum_{k=1}^{m} (x_k - y_k)^2 \right)^{1/2}$$

In some cases, defining similarities might be easier than dissimilarities, e.g. a similarity between two vectors can be defined as a correlation coefficient. For a

similarity value s between 0 and 1, the corresponding dissimilarity d can be found as: $d = 1 - s$ or $d = \sqrt{\{1-s\}}$. There are many other possibilities of defining a (dis)similarity measure for real-valued, discrete-valued, mixed-valued vectors, as well as sets. For further information, see e.g. [Cox, 1995; Everitt, 1997].
When the data contains features with different orders of magnitude (some features described by small numbers, others by relatively large ones), we may want to prevent too much contribution of a feature with large numbers while computing dissimilarities. Therefore, proper normalization is important. The most common way is to standardize the data, which means that all features are transformed such that they have mean 0 and variance 1. Another possibility might be to weigh each feature separately according to the prior knowledge on the feature's origin or characteristics.

Metric MDS-theory

Suppose a set of n objects with dissimilarities δ_{ij}, $i,j=1,2,\ldots n$, measured between all pairs of objects, is given. Our aim is to find a possibly low-dimensional space such that the distances between points are as close as possible to the corresponding dissimilarities. The dissimilarities can describe the relations between objects represented originally in a high-dimensional space or they can just be measured (e.g. road distances) or given (by human judgments). There are different ways of preserving the structure of the data, giving rise to somewhat different techniques of MDS.

Classical scaling

Classical scaling originated in 1930s when [Young, 1938] showed, how starting with a matrix of Euclidean distances between all pair of points, the coordinates for these points can be found in such a way that these distances are preserved. Then, this technique was used for projection. This method is popular, because it gives an analytical solution, which does not require an iterative algorithm.
Given a Euclidean distance matrix $\Delta \in \Re^{n \times n}$ between n objects, a distance preserving mapping onto an Euclidean space can be found. In other words, the dimensionality $k \leq n$ and the configuration $X \in \Re^{n \times k}$ (with objects represented as rows of X) can be found such that the (squared) Euclidean distances are preserved. Note that having determined one configuration, another one can be found by a rotation or a translation. To remove this last degree of freedom, without loss of generality, the mapping will be constructed such that the origin coincides with the centroid (i.e. the mean vector) of the configuration X. X can be defined based on the relation between Euclidean distances and inner products. It can be proven [Borg, 1997] that: $\Delta^{(2)} = \boldsymbol{b}\,\mathbf{1}^T + \mathbf{1}\,\boldsymbol{b}^T - 2B$, where $\Delta^{(2)}$ is a matrix of square Euclidean distances, B is the matrix of inner products of the underlying configuration X, i.e. $B = X X^T$ and $\boldsymbol{b}$ is a vector containing the diagonal elements of B. B can also be expressed as:

Formula 1

$$B = -\frac{1}{2} J \Delta^{(2)} J \quad \text{where} \quad J = I - \frac{1}{n} \mathbf{1}\mathbf{1}^T \in \Re^{n \times n}$$

and I is the identity matrix. J is the centering matrix, i.e. it projects the data such that the final configuration has zero mean. Then, the factorization of B by its eigendecomposition can be found as:

Formula 2

$$X\, X^T = B = Q\, \Lambda\, Q^T,$$

where Λ is a diagonal matrix with the diagonal consisting of the first non-negative eigenvalues [Borg, 1997], ranked in descending order, followed by the zero values, and Q is an orthogonal matrix of the corresponding eigenvectors. For $k \leq n$ non-zero eigenvalues, a k-dimensional representation X can then be found as:

Formula 3

$$X = Q_k \sqrt{\Lambda_k}, \quad Q_k \in \Re^{n \times k}, \quad \sqrt{\Lambda_k} \in \Re^{k \times k}$$

where Q_k is a matrix of the first k leading eigenvectors and $\sqrt{\{\Lambda_k\}}$ contains the square roots of the corresponding (largest) eigenvalues. Note that X, determined in this procedure, is unique up to rotation (the centroid is now fixed), since for any orthogonal matrix T, $X\,X^T = (XT)\,(XT)^T$.
X is constructed such that the features are sorted according to decreasing contribution while computing the Euclidean distance. Since dissimilarities are noisy measurements, k can be close to n, which means that a representation in a higher-dimensional space is found. If only q eigenvalues are relatively large, then only q dimensions can be considered, since the remaining ones can be treated as insignificant information. Then, the q-dimensional representation becomes $X_q = Q_q \sqrt{\{\Lambda_q\}}$, which means that insignificant dimensions have been disregarded. In such a case, distances are preserved only approximately. This is possible since in case of Euclidean distances, the solution provided by classical scaling is the same as the solution given by the Principal Component Analysis [Manly, 1994], provided that the original data is given in a high-dimensional space.
In the classical scaling method different symmetric dissimilarity measures can be used, i.e. Δ can be any symmetric dissimilarity matrix. However, when other types of distances are used, e.g. dissimilarities based on human judgment, negative eigenvalues may appear in formula 2 and X cannot be formally found, since the square root cannot be taken. There are a number of ways to approach this problem, where the easiest one is to neglect the negatives and take into account the positive eigenvalues only (see [Borg, 1997; Cox, 1995; Davidson, 1992]).

Sammon mapping

The classical scaling technique is a linear projection found as an analytical solution. The Sammon mapping [Sammon, 1969; Borg, 1997; Cox, 1995], on the other hand, is a non-linear projection realized via an iterative process. In such a process, a criterion is needed for deciding whether one configuration is better than another. For that purpose, an error function, called stress, is considered, which measures the difference between Euclidean distances of the present configuration of n points in an m-dimensional space and the original dissimilarities. Let Δ be the originally given dissimilarity matrix and D be the distance matrix for the projected configuration. The original Sammon stress function [Sammon, 1969] is defined as follows:

$$S = \frac{1}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\delta_{ij}} \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{\left(\delta_{ij}-d_{ij}\right)^2}{\delta_{ij}}$$

and yields in fact a badness-of-fit measure for the entire representation. In general, the stress function can be defined in a number of ways, for instance as:

$$S = \frac{1}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\delta_{ij}^{t+2}} \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\delta_{ij}^{t}\left(\delta_{ij}-d_{ij}\right)^2\right), \quad t = \ldots,-2,-1,0,1,2,\ldots$$

which results in the following measures:

$$S_{-2} = \frac{1}{(n-1)(n-2)} \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\frac{\delta_{ij}-d_{ij}}{\delta_{ij}}\right)^2$$

$$S_{-1} = E_S$$

Formula 4

$$S_0 = \frac{1}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\delta_{ij}^{2}} \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\delta_{ij}-d_{ij}\right)^2$$

$$S_1 = \frac{1}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\delta_{ij}^{3}} \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\delta_{ij}\left(\delta_{ij}-d_{ij}\right)^2\right)$$

$$S_2 = \frac{1}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\delta_{ij}^{4}} \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\delta_{ij}^{2}\left(\delta_{ij}-d_{ij}\right)^2\right)$$

These values are normalized in one way or another in order to avoid a scale dependency. Each of the stress functions mentioned above emphasizes a different aspect of the geometric relations between points, i.e. it emphasizes, to some extent, either small or large distance. For instance, S_{-2} emphasizes very small distances, while S_1 emphasizes the large ones. S_0 provides a nice balance between large and small distances. The same stress measures can also be applied for squared distances, which give rise to more global MDS maps.

The problem of finding the right configuration of points in a low-dimensional space resolves itself into an optimization problem. We are interested in obtaining such a configuration that the stress function is minimum. Each of the presented stress functions is optimal, when all the original dissimilarities δ_{ij} are equal to d_{ij}, the distances between the projected points. However, this is not likely to happen. Therefore, the found distances will be imperfect, distorted representations of the relations within the data. The larger the stress, the greater the distortion.

To find an MDS representation, we start from the initial configuration of points $\{\boldsymbol{x}_i\}_{i=1}^{n}$ (e.g. randomly chosen) for which all the pair-wise distances are computed and the specified stress value is calculated. Next, the points are adjusted so that the stress will decrease. In other words, we try to improve this configuration by shifting around all points in small amounts to approximate better and better the ideal model relation $d_{ij} = \delta_{ij}$ for $i,j=1,2,\ldots,n$. This is done in an iterative manner, until a configuration corresponding to a minimum of the stress is found. In such a procedure, a steepest descent or pseudo-Newton algorithm or iterative majorization [Borg, 1997, Cox, 1995] can be used to search for the minimum of the stress function. It is important to emphasize that usually a local minimum is found.

Artificial data

An example of the potential of a non-linear mapping is given for an artificial dataset, representing points lying on two circles in a 3-dimensional space, both with radius equal to 1.0. The circles are placed on two planes parallel to the *yz*-plane, with the distance 1. The data is shown in Figure 6. The linear and non-linear MDS projections on a plane are shown in Figure 7.

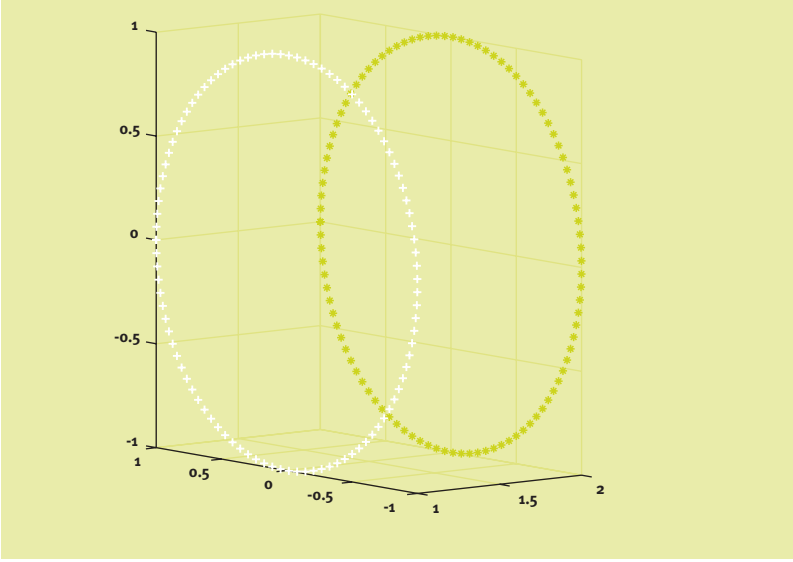


Figure 6
Two circles.

In the case of classical scaling (linear projection), each two corresponding points from both circles are mapped onto a single point. It looks, therefore, as if the data comes from *one* circle. In this way, some important information is lost, namely the existence of the second circle. The non-linear mapping illustrates two oval, closed curves. Of course, it might be difficult to judge if those two shapes represent symmetrical bends in the original space, but that is the price to be paid for a projection on a non-linear subspace. In general, non-linear mappings reveal more 'hidden' structure of the data.

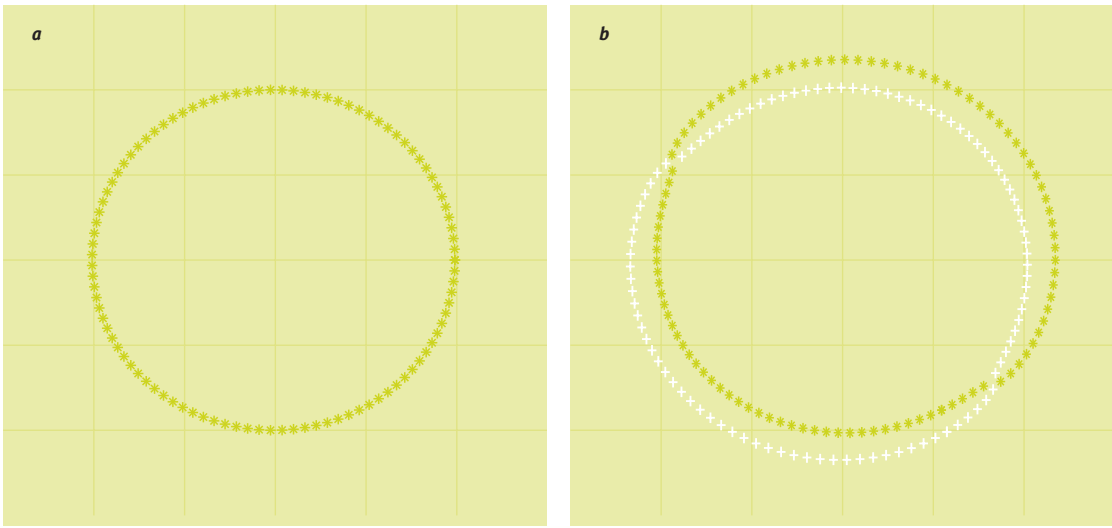
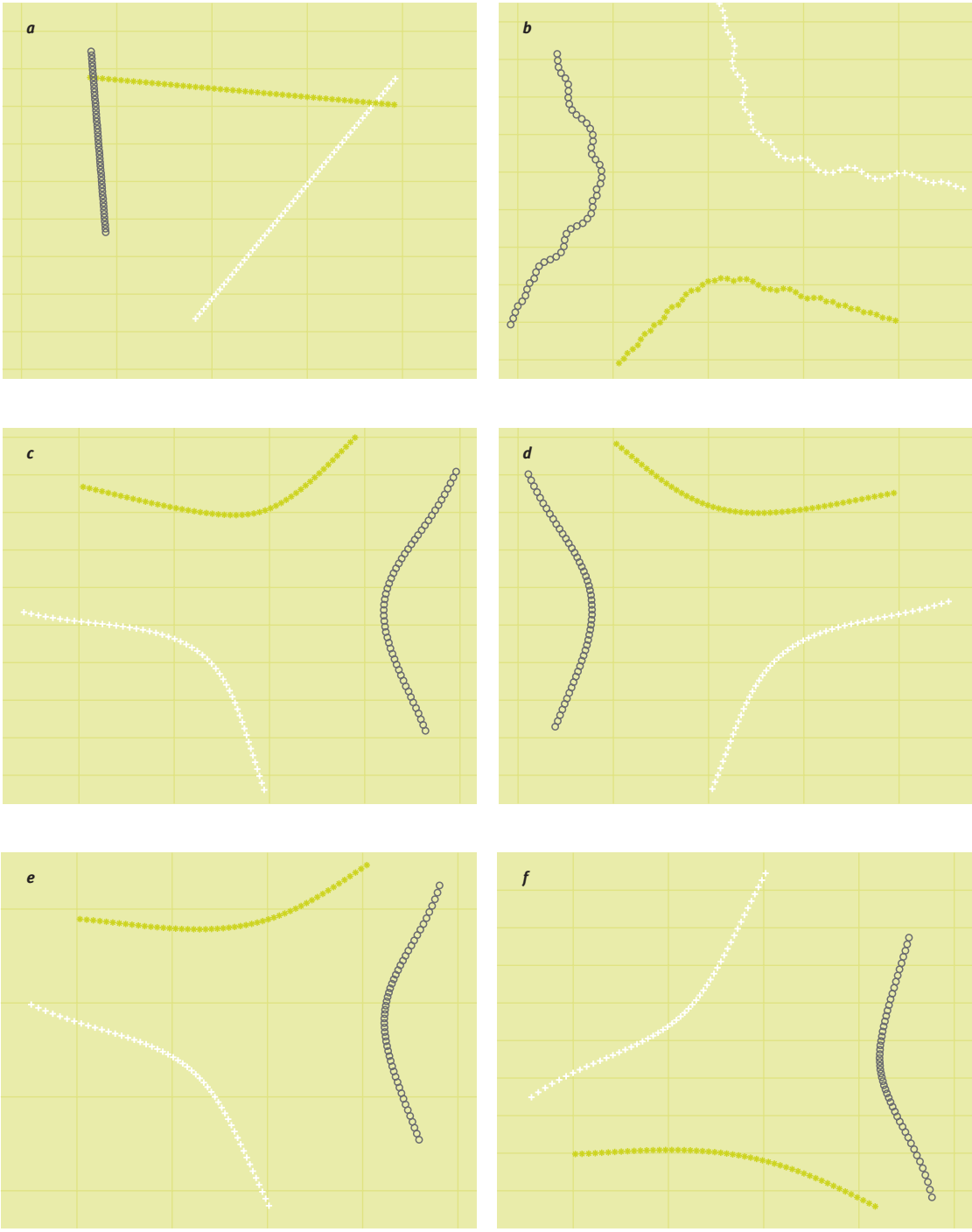


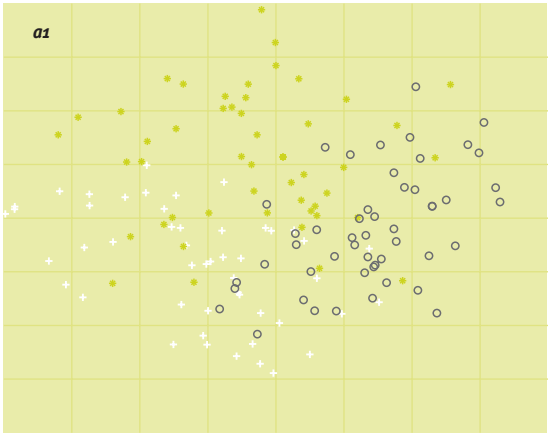
Figure 7
The MDS maps of two circles in a 3-dimensional space.
a *classical scaling;*
b *Sammon mapping.*

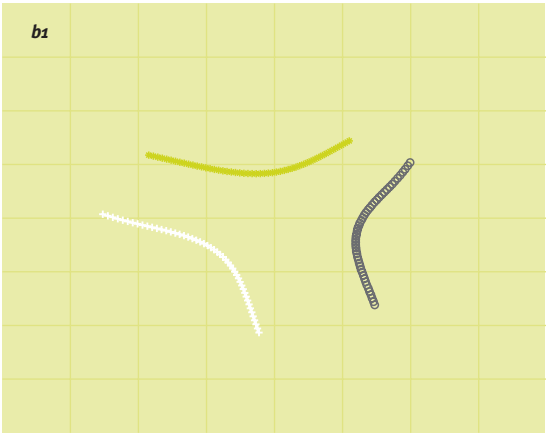
Figure 8 (opposite page)
The MDS maps of the three non-crossing lines in 5D.
a *classical scaling;*
b *Sammon mapping S_{-2};*
c *Sammon mapping S_{-1};*
d *Sammon mapping S_0;*
e *Sammon mapping S_1;*
f *Sammon mapping S_2.*

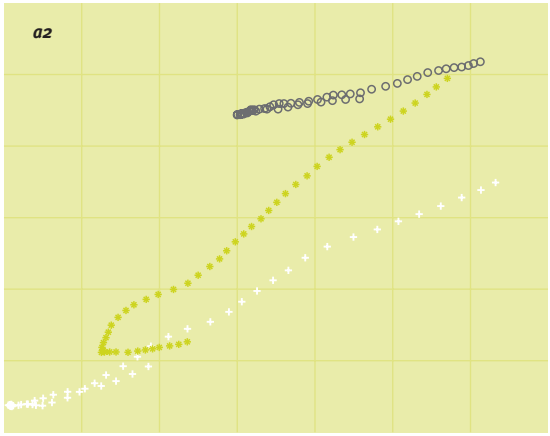
To illustrate the differences between the stress measures, as well as in non-linearity of the projections involved, an artificial example of points lying on three non-crossing, non-parallel lines in a 5-dimensional space is considered. Figure 8 presents 2-dimensional MDS maps constructed by the optimization process of the stresses (formula 4). The observation of the mapping results confirms the potential superiority of non-linear mappings over linear projection. On the basis of classical scaling, one can draw the false conclusion that the data represents three crossing lines in a higher-dimensional space, while the Sammon mapping result suggests that the data consists of three *non-crossing* curves, but of course, not necessarily lines. Therefore, linear and non-linear mappings are useful while used together, i.e. the methods complement each other.
The Sammon mappings are ordered with respect to the non-linearity involved in projections. By minimizing S_{-2}, one focuses on preserving very small distances, by which all the perturbations are visible (Figure 8b). By optimizing S_2, on the contrary, one tries to preserve mostly large distances, and as a result, the curves start to resemble straight lines in some ways. The stress S_0, keeps the balance between preserving small and large distances. The choice of a stress function depends on which geometric properties a MDS map should have.

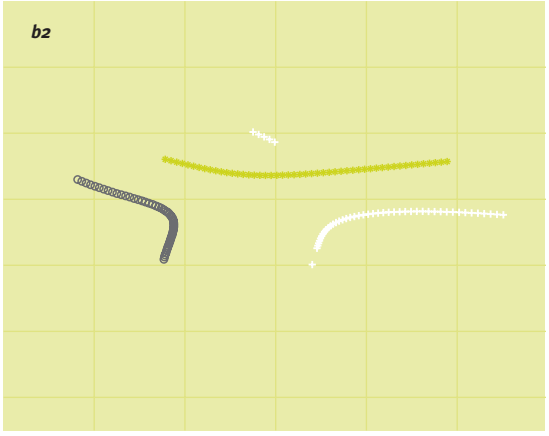


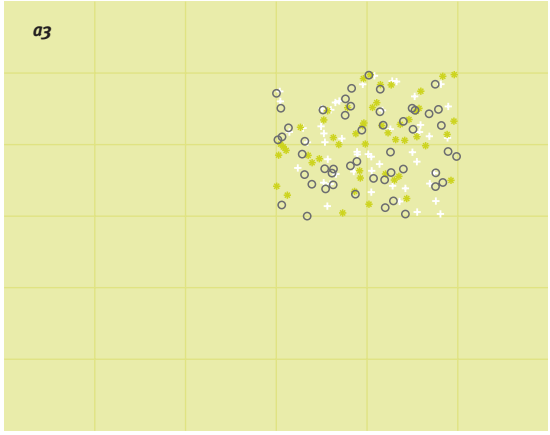
When no preferences are given, our experience suggests that either S_0 or S_{-1} can be used.

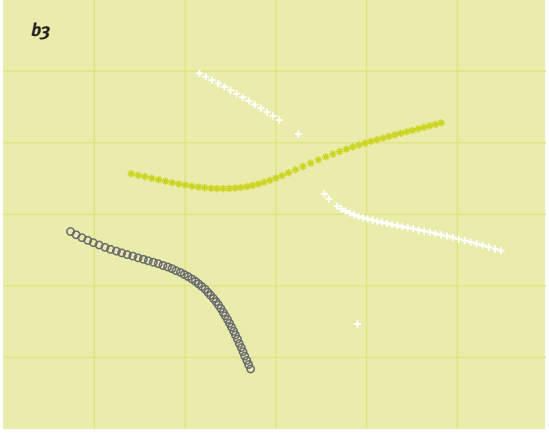












To illustrate that for the Sammon mapping, an initial configuration might in some cases be crucial for the final result, the Sammon projection with the stress S_0 has been used. Figure 9 presents the differences between the final configura-

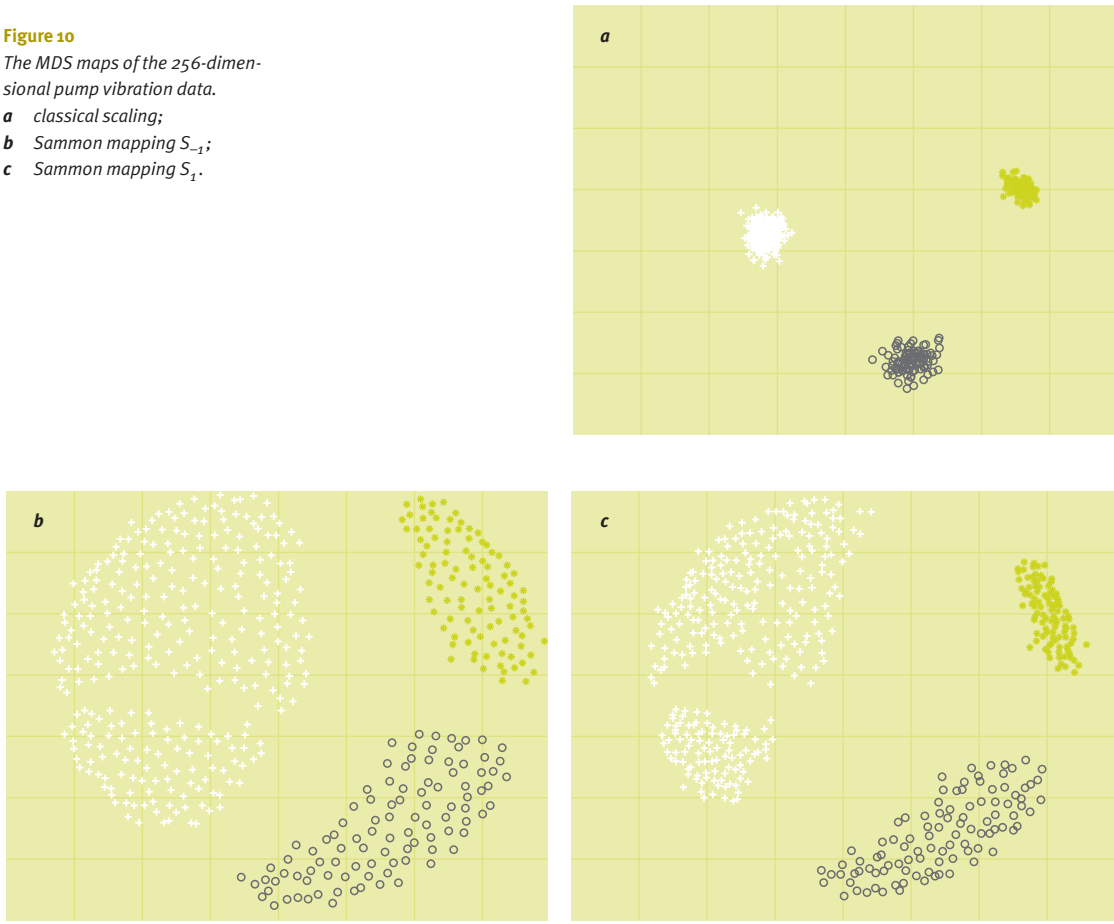
Figure 9 (opposite page)
Sammon maps of three non-crossing lines in 5D: the initial and final configurations.
a *initial maps;*
b *corresponding final maps with the stress* E_{S-2}.

tions, when started from different initial configurations. It seems, from our experience, that initializing the Sammon projection by classical scaling often gives good results. Another advantage is that the minimization process is also relatively short. Therefore, such initialization was applied in most cases. It is, however, always useful, to see the MDS result based on random initialization. The optimization procedure initialized by the classical scaling may, in some cases, stack easily in a first local minimum.

Pump data

Here we consider a real example. Vibration was measured with three accelerometers mounted on a submersible pump which operated in three states: normal, presence of imbalance and presence of bearing failure. Moreover, the bearing failure was measured at three different operating speeds. The data consists of 500 observations with 256 spectral features of the acceleration spectrum (see [Ligteringen, 1997]). It is known [Ypma, 1997] that the data has a low intrinsic dimensionality and that it probably lies in a non-linear subspace of a 256-

Figure 10
The MDS maps of the 256-dimensional pump vibration data.
a *classical scaling;*
b *Sammon mapping* S_{-1};
c *Sammon mapping* S_1.



dimensional space. The MDS projections based on the Euclidean distances are shown in Figure 10. Classical scaling reveals three non-overlapping classes. A non-linear projection, here the Sammon mapping with the stresses S_{-1}, and S_1 show, however, much more structure in the data. By the Sammon results, additional information can be gained: that the class of bearing failure (white class) is composed of two or three subclasses, which corresponds to the fact that three operating speeds were used.

If this data was used for the dimension reduction purposes, basically, in order to judge the intrinsic dimensionality, the MDS mapping should be performed to a number of dimensions. Then, the plot of the stress as a function of dimensionality can be obtained, as in Figure 11. From such a figure, one can find a potential intrinsic dimensionality, which corresponds to a point where the rapid decrease in the stress function stops. In our case, it would be 6–7.

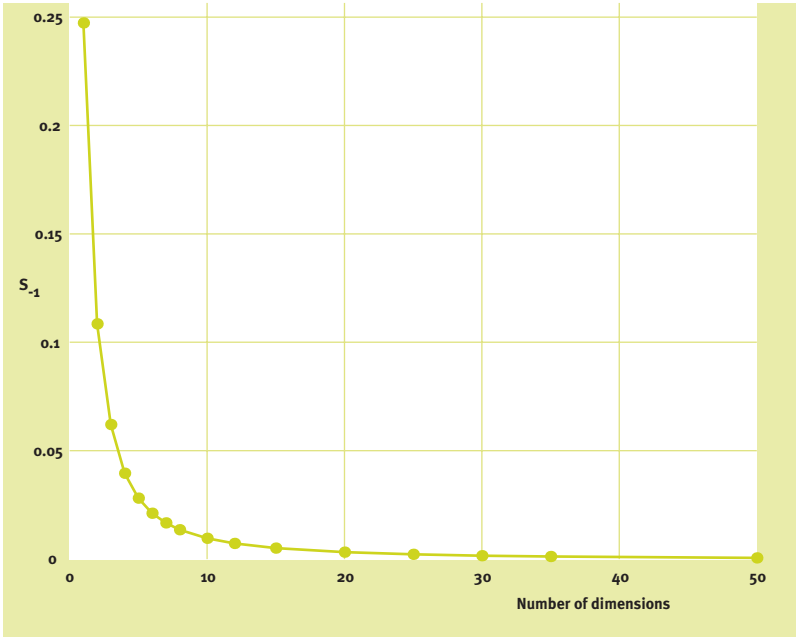


Figure 11
S_{-1} versus the dimensionality.

Non-metric MDS - Kruskal mapping

Kruskal mapping [Kruskal 1978; Cox, 1995] is used for non-metric MDS. Again, pair-wise dissimilarities δ_{ij} are given for n objects. A configuration of points is sought in an m-dimensional space such that the distances d_{ij} between points in a lower-dimensional space match as well as possible the original dissimilarities δ_{ij}. Here, an assumption is made that the dissimilarities are transformed by some monotonic, increasing function and that they might be represented by the distances in an m-dimensional space. This makes it different from the metric MDS methods. Once the dissimilarities and the function of transformation are

chosen, this problem becomes one of finding an appropriate algorithm for minimizing the stress function.

Let Δ be the dissimilarity matrix originally given. We start from the initial configuration of points $\{\boldsymbol{x}_i\}_{i=1}^{n}$ (i.e. the matrix $X \in \Re^{n \times m}$) and calculate the distance matrix D. The starting configuration may e.g. consist of randomly chosen points. For each evaluated distance d_{ij}, a regression on the original dissimilarity δ_{ij} for $i,j=1,\ldots,n$ is made. The regression can be linear, polynomial or monotonic. For example, a linear regression assumes that

$$d_{ij} = a + b\,\delta_{ij} + e_{ij} \quad ,$$

where a and b are constants and e_{ij} is an error.

Let us denote by $\hat{d}_{ij}$ (called disparities) the distances obtained from the regression equation, i.e. $\hat{d}_{ij} = a + b\,\delta_{ij}$. The goodness-of-fit between distances and disparities is then measured by the Kruskal stress:

$$K = \sqrt{\frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_{ij}^2}}$$

The coordinates of each point are changed so that the stress can be reduced. For the next configuration of points again the distance matrix D is computed, serving for the regression equation on the dissimilarities and the whole procedure is repeated. This is performed iteratively many times in order to minimize the stress. For the relation $\hat{d}_{ij} = \delta_{ij}$, the whole problem reduces to the Sammon mapping with the stress S_o.

Summary

Multidimensional scaling refers to a group of linear and non-linear projection methods from the area of exploratory data analysis. These methods facilitate data visualization and exploration. They can also be used for dimension reduction. These projection techniques aim to preserve all pair-wise, symmetric dissimilarities between data objects, resulting in a faithful, low-dimensional representation of the geometrical relations between the points. Such a configuration of objects is usually found in a Euclidean space, although any other L_p space can be considered, in general. Non-linear methods, namely Kruskal mapping or the variants of Sammon mapping, can reveal more structure and cluster tendencies than the linear projection, i.e. classical scaling. They are, however, much more time-consuming. To understand the data better, both linear and non-linear methods should be used, since they complement each other.

Another purpose of such projection techniques is to find rules that would explain observed dissimilarities, especially these obtained by human judgment,

and would help to describe the data structure in simple terms, which is of importance in marketing, sociology or psychology.
In summary, the multidimensional scaling methods provide a tool for a better comprehension of the data.

References

- Borg, P. Groenen. (1997). Modern Multidimensional Scaling. Springer Verlag, Berlin
- Cox, T.F., M.A.A. Cox. (1994). Multidimensional Scaling. Chapman and Hall, London
- Davison, M.L. (1983). Multidimensional Scaling. Krieger, Malabar, Florida
- Everitt, B.S., S. Rabe-Hesketh. (1997). The Analysis of Proximity Data. Arnold, London
- http://www.psychology.adelaide.edu.au/members/staff/michaellee.html
- Kruskal, J.B., M. Wish. (1978). Multidimensional Scaling. Sage Publications, Newbury Park, CA
- Ligteringen, R., R.P.W. Duin, E.E.E. Frietman, A. Ypma. (1997). Machine Diagnostics by Neural Networks, Experimental Setup. In: H.E. Bal, H. Corporaal, P.P. Jonker, J.F.M. Tonino. (eds.). ASCI'97 Proceedings 3rd Annual Conference of the Advanced School for Computing and Imaging (Heijen, NL, June 2-4). ASCI, Delft. pp185-190
- Manly, B.F.J. (1994). Multivariate Statistical Methods, a Primer. Chapman & Hall, London
- Sammon, J.W. (1969). A Non-Linear Mapping for Data Structure Analysis. Trans. Comp **C-18**:401-409
- Young, G., A.S. Householder. (1938). Discussion of a Set of Points in Terms of their Mutual Distances. Psychometrika **3**:19-22
- Ypma, R. Ligteringen, E.E.E. Frietman, R.P.W. Duin. (1997). Recognition of Bearing Failures Using Wavelets and Neural Networks. 2nd UK Symposium on Applications of Time-Frequency and Time-Scale Methods. University of Warwick, Coventry, UK. pp69-72

6.2.6 Clustering

Ad Feelders[1]

Introduction

The objective of clustering is to put objects (persons, households, transactions, and so on) into a number of groups in such a way that the objects within the same group are similar, but the groups are dissimilar.
Each object is described by a number of variables (also called features or attributes). The similarity between objects, or between an object and a group, is determined on the basis of this description. The measurement of similarity is crucial to many clustering methods, and the proper measure of similarity depends on the application as well as the types of variables involved.
Many techniques have been developed to cluster objects into groups. We discuss some of the most popular ones:

- Hierarchical clustering.
- Partitioning methods.
- Model-based clustering.

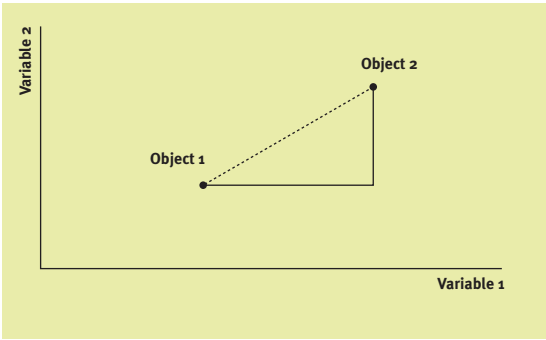
Hierarchical clustering and partitioning methods work on a dissimilarity matrix that contains a measure of the dissimilarity between each pair of objects. In model-based clustering one assumes that each cluster can be described by a probability model. This brings the clustering problem within the realm of statistical inference.

Dissimilarity measures

Many clustering methods work on a dissimilarity matrix that contains a measure of the dissimilarity between each pair of objects in the data set. In this section we look at different dissimilarity measures. The appropriate dissimilarity measure depends on the type of variables that describe the objects, and on the application. For numeric data common choices are Euclidian distance and Manhattan (city block) distance. To determine the Manhattan distance, we simply take the distance between the two objects measured on each variable, and sum over all variables. For the two-dimensional case this is illustrated in Figure 1. The Manhattan distance corresponds to the length of the solid line in this picture. The Euclidian distance corresponds to the length of a straight line between the two objects (the dashed line in Figure 1).

Finding an appropriate dissimilarity measure when some variables are numeric and some are discrete is somewhat more difficult. Whatever dissimilarity measure we decide to use for the data at hand, once we have computed the dissimilarity between each pair of objects we can start clustering.

1 Dr A.J. Feelders, ad@cs.uu.nl, Utrecht University, Institute of Information & Computing Sciences, Utrecht, The Netherlands



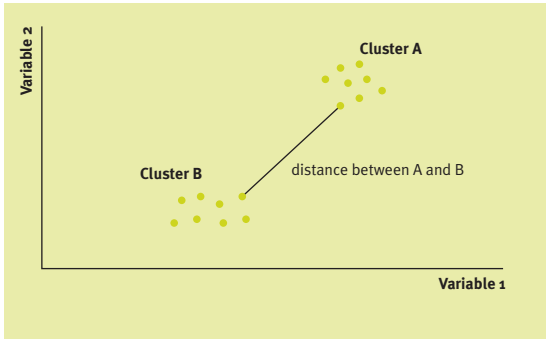


Figure 1 (left)
The Euclidian distance (dashed line) and the Manhattan distance (solid line) between object 1 and 2.

Figure 2 (right)
Single-linkage distance between cluster A and B, based on the Euclidian distance.

Hierarchical clustering

Now we have discussed several ways to compute the dissimilarity between pairs of objects, we turn to the use of these dissimilarities to determine a clustering of the objects.

Hierarchical clustering is one of the most popular clustering techniques in use. It comes in two flavors: agglomerative and divisive. We confine our discussion to the more popular agglomerative version. In agglomerative clustering we start with each data point (object) as a separate cluster, and we successively merge the two closets clusters until finally we reach a single cluster which contains all data points. We can define different measures for the closeness of two clusters leading to different variations of agglomerative clustering.

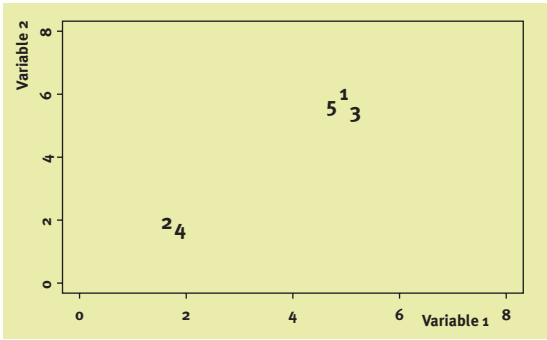
In Figure 2 the length of the solid line indicates the single-linkage distance between clusters *A* and *B*, based on the Euclidian distance as a dissimilarity measure. Single linkage simply looks at the distance between all pairs of points, one from each cluster, and takes the smallest of those distances to be the distance between the two clusters.

Let's look at a small example to illustrate the process of agglomerative clustering. We start with a small data set with 2 measurements made on 5 objects (see Table 1).

Table 1
Five objects measured on two variables.

	Variable 1	**Variable 2**
Object 1	5.0	6.2
Object 2	1.7	2.1
Object 3	5.2	5.7
Object 4	1.9	2.0
Object 5	4.8	5.9

Since we only have 2 dimensions, we can make a simple plot of the data (see Figure 2), which allows us to make a clustering just by visual inspection of the plot. We clearly see 2 distinct clusters, one cluster containing objects 2 and 4



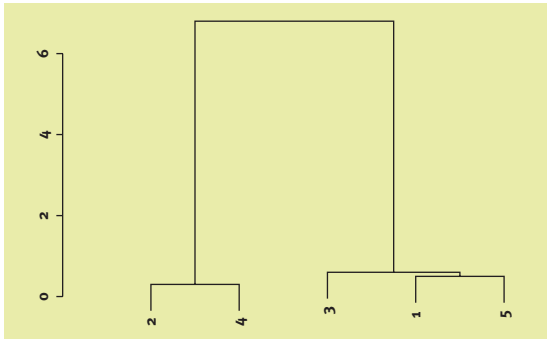


Figure 3 (left)
Visual display of the 5 objects from Table 1.

Figure 4 (right)
Dendrogram displaying hierarchical clustering of objects in Table 1.

around the point (2,2) and the other cluster containing objects 1,3 and 5, around the point (5,6).

Initially, each data point forms a cluster of its own and we simply look for the two objects with the smallest distance to find the two clusters to be merged. Objects 4 and 2 are most similar, since their distance $|1.7–1.9|+|2.1–2.0|=0.2+0.1=0.3$, is smaller than the distance between any two other objects. This means we make a new cluster {4,2}, and the single-point clusters {4} and {2} are removed. Now we have to compute the distance between the newly formed cluster, and the remaining clusters {1}, {3} and {5}. If we use single linkage to compute the distance between clusters, the distance between {1} and {2,4} becomes the minimum of the distance between 1 and 2, and the distance between 1 and 4. The distance between 1 and 4 (7.3) is the smallest of the two, so this becomes the distance between the two clusters. After the distances have been updated in this way, we again look for the two closest clusters and merge them, until finally we only have one cluster left.

Figure 4 displays a dendrogram of a hierarchical clustering of the objects in Table 1, based on Manhattan distance between objects and single-linkage distance between clusters. At the bottom of the dendrogram, the individual objects are displayed as the initial clusters. The dissimilarity between clusters that are merged can be read off from the vertical scale at the left. For example, the first clusters to be merged are {2} and {4}, and their dissimilarity is 0.3. The final merge of the clusters {2,4} and {1,3,5} shows a large 'jump' in dissimilarity, suggesting that these two clusters provide a good representation of the structure in the data.

Partitioning methods

Partitioning methods search directly for a division of the objects into a number of groups that maximizes a quality criterion which reflects what we consider to be a good clustering.

To find the best partitioning is computationally quite complex, because the number of distinct partitions of the objects into different groups rapidly gets

very large. For example, if we have 100 data points and we want to partition them into 5 groups, then the number of distinct partitions is 10^{68}. Therefore it is infeasible to simply check all partitions and select the best one. Usually a hill-climbing algorithm is used to find a good grouping.
A typical hill-climbing approach to this problem roughly looks like this:

1 Find some initial partition into the required number (k) of groups.
2 Calculate the change in quality by moving each object from its own to another cluster.
3 Make the change which leads to the greatest improvement in quality.
4 Repeat steps 2 and 3 until no movement of a single object causes the quality to improve.

There are many possibilities for measuring the quality of a partition. In case of numeric data, one can use for example the sum of squared Euclidian distances of all objects to the mean of the cluster to which they are assigned. A hill-climbing algorithm that uses this quality criterion is the k-means algorithm [Hartigan, 1975].
So far we have assumed that the required number of groups is known in advance to the analyst. This is in fact rarely the case in practical data analysis: usually the analyst has to estimate the appropriate number of groups from the data as well. As remarked by [Everitt, 1993], most proposed methods are relatively informal and involve plotting the quality criterion against the number of groups. We plot the value of the quality criterion for k=1,2,3,… and look for large jumps to determine the appropriate number of groups.

Model-based clustering

The clustering methods considered so far use plausible dissimilarity measures, and quality criteria for finding a good partitioning. They leave us in the dark, however, when it comes to questions concerning the confidence we may have in the solution obtained.
In model-based clustering the data to be analyzed is viewed as a sample from a population that consists of a number of subpopulations (clusters or components), where each subpopulation can be described by a probability model. By using the theory of probability and statistical inference, we can start to look at questions such as:

– What is the *probability* that an object belongs to a particular cluster?
– What is the most likely number of clusters present in the population?

Of course the answers obtained to those questions always depend on the modeling assumptions we make along the way. We typically assume that each cluster is a member of the same parametric family of probability distributions. Examples of useful distributions are:

– Numeric data: multivariate normal distribution.
– Binary data: multivariate Bernoulli distribution.

Such a 'weighted average' of a number of component distributions is called a mixture distribution [McLachlan, 1988; Everitt 1981]. Estimation of the parameters of the statistical model is usually quite complex involving iterative techniques such as expectation maximization (EM).
Once the parameters of the statistical model have been estimated, we can compute the probability of cluster membership for each observation. So rather than just an outright assignment to one cluster (as is the case e.g. in partitioning methods), we obtain probabilities for each cluster. Of course, when an assignment to one cluster is required, we can simply pick the cluster with highest probability.
Another important issue is choosing the number of components. So far, we have assumed that we fix this number in advance. In many cases we don't know the number of clusters, however, so we would like to estimate it from the data as well. Statistical tests and other measures have been developed to obtain an estimate of the number of clusters present in the data.

Summary

Cluster analysis is concerned with finding groups of similar objects in data. Most clustering algorithms work with the notion of distance or dissimilarity between objects. Finding an appropriate dissimilarity measure for the application and type of data concerned is one of the difficult elements of cluster analysis. Once the dissimilarity between each pair of objects has been computed, the actual clustering can proceed. In agglomerative hierarchical clustering we start with each object as a separate cluster, and successively merge the two closest clusters until we finally reach one cluster that contains all the data points. Large jumps in dissimilarity are taken to be suggestive for the number of clusters present in the data.
Alternatively, we can use partitioning methods where we fix the number of clusters k in advance. Partitioning methods try to find a division of the data points into k groups that optimizes some criterion that is indicative of the quality of the grouping. Since exhaustive search is intractable, one usually applies hill-climbing to find a reasonably good solution. To determine the appropriate number of groups, one usually tries different values for k and looks for large changes in the value of the quality criterion.
Model-based clustering is a different approach, based on probability theory and statistical inference. One assumes that each cluster can be modeled by a probability distribution, and the relevant parameters are estimated using, for example, maximum likelihood. Model estimation is typically quite complex, and requires iterative techniques such as expectation maximization. The advantage

of model-based clustering is that modeling assumptions are made explicit (through the specification of the probability model) and it provides us with probabilities that each object belongs to a particular cluster, rather than just an outright assignment. Furthermore, we can use statistical testing to estimate the number of clusters present in the data.

References and further reading

The CD-rom shipped with this book contains a more elaborate discussion on clustering.

- Everitt, B.S., D.J. Hand. (1981). Finite Mixture Distributions. Chapman and Hall, London
- Everitt, B.S. (1993). Cluster Analysis. Third edition. Edward Arnold, London
- Hartigan, J.A. (1975). Clustering Algorithms. Wiley, New York
- McLachlan, G.J., K.E. Basford. (1988). Mixture Models, Inference and Applications to Clustering. Marcel Dekker, New York

6.2.7 Classification/Decision tree learning

Maarten van Someren[1]

This method is characterized by the type of model that is constructed: a decision tree. Suppose that the training data consist of values on variables and a class label for each example. Within a decision tree model, the nodes in the tree are associated with tests on the variables. With the outcomes of tests a branch is associated to the next node. The leaves of the tree are associated with class labels. Not all variables that appear in the training data need to appear in the tree. The most common type of tests are equality of variables (e.g. color = red) and intervals for numerical variables (e.g. 10" < size < 25"). Sub trees can be different and in this way a decision tree can express non-linear effects. For example, if color = red, then the interval 10-25 may predict the class A, but if color = yellow, interval 30-50 may predict A.
A simple example of a decision tree for patient diagnosis is given in Figure 1.

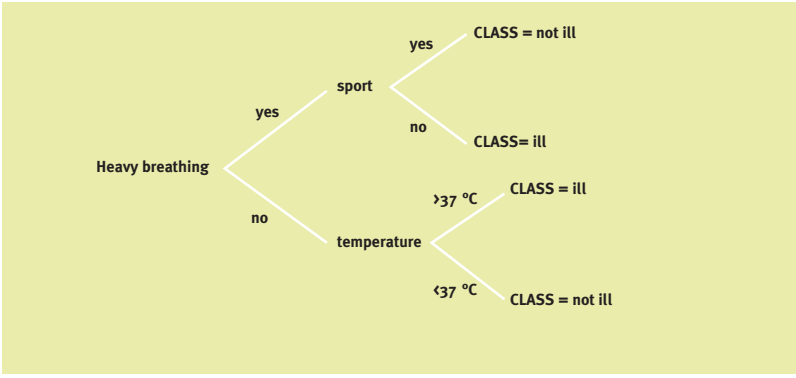


Figure 1
The nodes of the tree are 'heavy breathing', 'sport' and 'temperature'. The leaves are CLASS= ill and CLASS = not ill. The tree is interpreted from left to right. For example, if for a patient 'heavy-breathing' is 'yes' and 'sport' is 'no' then CLASS is 'ill'.

The algorithm for constructing a decision tree for a dataset works 'top down': In each step a test for the actual node is chosen (starting with the root node; decision trees are special trees that have their root at the top!), which best separates the given examples by their classes or in other words which best predicts the class. The quality of a test is measured by the impurity/variance of example subsets. The most common measure is based on a measure for the amount of information: the sum of the amount of information that remains in the subsets created by splitting the data according to the test on a variable, weighted by the proportion of examples in each subset. The best splitting (that is, the best predicting) test is selected, the data are split accordingly. The procedure is applied recursively to all subsets until they all belong to a single class.
in the machine learning literature the term TDIDT for Top-Down Induction of Decision Trees is often used. The term 'classification tree' is also used for what we call decision trees here.

1 Dr M. van Someren, maarten@swi.psy.uva.nl, The Universiteit van Amsterdam, Faculty of Social and Behavioural Sciences, Department of Psychology, Amsterdam, The Netherlands

Table 1
Top-Down Induction of Decision Trees (TDIDT) procedure.

```
TDIDT(Data, Tree):
IF   sameClass(Data, Class) THEN
     Tree := leaf(Class)
ELSE
find best attribute(Data, BestAttribute)
splitData(Data, Attribute, SubDataSets)
addAttributeToTree(Attribute, Tree, NewTree)
Associate(SubDataSets, NewTree)
FOR EACH pair SubDataSet / Leaf in NewTree DO TDIDT(SubDataSet, Leaf)
```

Note: In economics and decision making, the term 'decision tree' is also used to express 'conditional plans' that involve external tests, have actions on the branches and a result (or pay off) at the leaves. For example

```
IF   clouds are grey
THEN take raincoat; start car and drive to bar (value = 50)
     sky is clear THEN take bike and go to beach
```

This a different notion of a decision tree that will not be discussed here.

TDIDT example

Suppose that we have a database with descriptions of patients. Medical examination has shown for these patients, if they had an urgent heart or lung problem or not. Each patient is described by four variables:

fever: no/some/high
sex: male/female
heavy breathing: yes/no
sported: yes/no
class: possible lung/heart problem: yes/no

Suppose that we have data on 8 patients (see Table 2).

The TDIDT algorithm now first evaluates how well each variable predicts class membership, that is, how 'pure' the distribution of classes is within values of the variable. Compare fever and sex. For fever we get the following distributions (See table 3). For sex we get Table 4.

Table 2
An example database with descriptions of patients.

	Fever	Sex	Heavy breathing	Sport	Class
1	No	male	yes	no	yes
2	High	male	yes	yes	no
3	No	female	yes	no	yes
4	High	female	yes	no	no
5	Some	male	no	yes	no
6	High	female	yes	no	yes
7	Some	female	yes	no	yes
8	High	male	no	no	no

Table 3
Fever distributions table.

Fever	Class = yes	Class = no
no	2	0
high	3	1
some	1	1

Table 4
Sex distributions table.

Sex	Class = yes	Class = no
male	1	3
female	3	1

Which predicts better? A simple (and acceptable) criterion is the number of errors, assuming that we predict the most frequent class. E.g. if fever=no, we predict class=yes and make no error. If fever is high, we predict class=yes and make 1 error. In this way, with fever we would make 2 errors and with sex also 2, so they are equally good. However, fever has more values and thereby splits the data into smaller subsets. Within each subset the distribution will on average be purer. In this case we should prefer sex over fever.

More on decision trees

The amount of information in a set with two classes 'pos' and 'neg' is defined as:

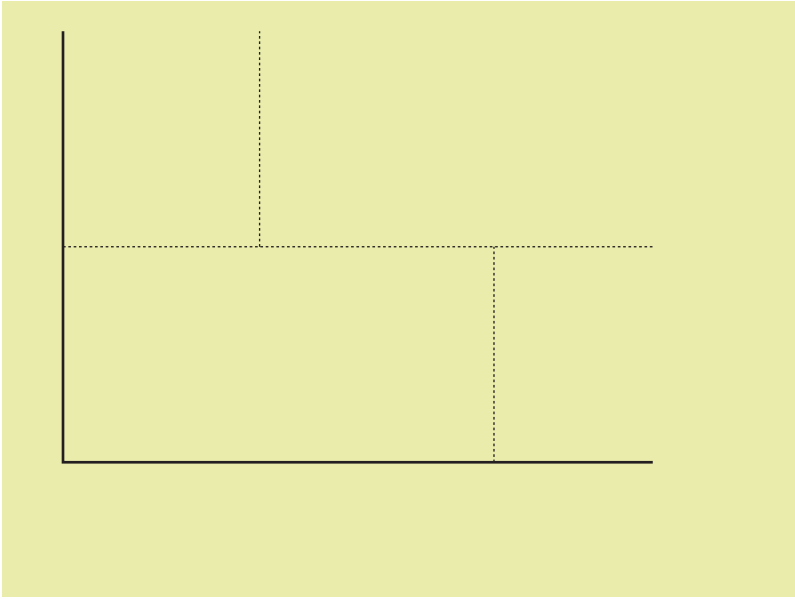
^{2}log proportion(pos)

If we have N subsets the total amount of information is the information in each subset, weighed by the proportion of examples in the subset, so we get:

Σ proportion$_i$ information$_i$

In geometric terms at each node in the tree the domain is split 'orthogonal' to the variable used at that node. Each subspace that is created here is further divided in the same way. This results in 'nested hyper rectangles'. Below we have two variables along the axes. First the space is split by the vertical axes and then the upper and lower space are split by the horizontal variable, at different points.

Figure 2
Orthogonal splits for two variables.



Strengths of this method are the simplicity and efficiency of the algorithm, the ability to discard irrelevant variables ('feature selection') and the readability of the resulting models. The main weaknesses are its limited ability to represent and find numerical models and the 'fragmentation' of data, leaving ever fewer data to learn sub trees.

TDIDT methods

There is a wide variety of methods that all use the TDIDT principle. We briefly review the most important variations.

Evaluation function

Using the 'purity' or homogeneity of subsets that result from splitting creates a bias in favor of variables and tests with many possible values. This is a problem, if the number of possible values varies widely. Variables with many possible test values will be a priori more homogeneous, but this advantage does not represent 'purity' in the domain and will lead to errors. Several corrections have been proposed that take the a priori expectation into account in estimating 'purity'.

Pruning

Splitting the data has the effect that for decisions deeper in the tree, ever fewer data are available. A statistical criterion is needed to decide, if there is enough evidence for a (further) split. A common method is the chi-square test which evaluates if the distribution after splitting is significantly different from the 'unsplit' distribution. Another approach is based on confidence intervals.

Discretization

Including a numerical variable in a decision tree requires the construction of a test on the values. Normally this is done by constructing intervals. Finding the number of intervals and the boundaries is a problem in itself. The most used method finds one boundary at a time and evaluates boundaries in the same way as other tests (using information gain and a statistical test). For the subsets corresponding to the two intervals a tree is constructed recursively and here the same numerical variable can be used again. A variation finds all candidate intervals once, for all data. This avoids much computation and also the problem of reduced amounts of data after splitting, but loses the possibility of creating different intervals in different sub trees.

Incremental learning

If training data come in one by one, a different algorithm must be used to update trees. Several methods have been proposed for this. To improve over the simple approach of storing all training examples and relearning from time to time, the idea is to maintain data about distributions at nodes in the tree and if the distribution changes, restructuring the tree: if the predictive power of a variable decreases, then it is shifted down in the tree or removed altogether.

Special trees

A number of extensions to the basic TDIDT scheme have been introduced. Many extensions use special tests in the tree. Simple extensions are the use of negation ('color IS NOT red') or disjunctions of values ('color = red OR yellow OR green'). A more complex extension is the use of relational descriptions of training examples and variables in tests (e.g. 'temperature(Day1, T1), temperature(Day2, T2), before(Day1, day2), above(T2, T1)').

Tree-based stacking

The tests that are used in a decision tree can be even more complex than those mentioned under 'special trees'. A test can be a discriminant function or a neural network of a special numerical function. Such tests can be constructed in a first pass on the data and then used as variables for decision tree learning. This way of using one method to construct models that are used as variables in another method is called 'stacking'.

References, links and tools

Ross Quinlan probably did most to promote the application of decision tree methods and to stimulate extensions and theoretical analysis. Statistical toolkits like SAS and SPSS include systems that perform TDIDT. Most data mining companies use it. Well-known are a popular shareware version (C5.0, distributed by the company Rulequest) and CART. All textbooks on machine learning and data mining include a discussion of decision tree methods, e.g. [Mitchell, 1997; Langley,1996; Weiss, 1998].

- Langley, P. (1996). Elements of Machine Learning. Morgan Kaufmann, San Francisco
- Mitchell, T.M. (1997). Machine Learning. McGraw-Hill, New York
- Weiss, S.M., N. Indurkhya. (1998). Predictive Data Mining. Morgan Kaufmann, San Francisco

6.2.8 Neural networks for data mining

Walter Kosters[1]

In many application areas neural networks are known to be valuable tools. This also holds for data mining. In this chapter we discuss the use of neural networks, we shall give an informal description of the way they work internally (just for one distinctive member of the large family of neural networks), and we finally focus on their usefulness for data mining. In particular we shall not deal with biological or psychological backgrounds. We only notice that the idea of neural networks originates from the physiology of the human brain.

General background

Neural networks are powerful general purpose learning devices. This sentence shows the strength (the general purpose character), which is also its weakness: its generality. Another weakness is the complex internal structure, which — if one finally understands the algorithms involved — still shows black box behavior: it is very hard to get an idea of the meaning of the internal computations.
Neural networks perform well in pattern recognition tasks, such as recognition of handwritten characters or spoken text. It should be noted that the way these networks learn (their 'training') is often supervised: the user should provide as many positive examples as possible; negative examples are also helpful. So for classification and clustering it is necessary to have classified or clustered input available. Note, however, that special neural networks are available that can cope with unsupervised situations.
It is easy to build a neural network that tries to solve a given problem. In fact, many software packages contain plug and play neural networks. But still many parameters need to be set, the training stage may take a while, and the tuning therefore can be awkward. State of the art hardware is a prerogative.
Another feature of neural networks is their random behavior. This does not mean that they act or react randomly, but that their training process contains random elements. For instance, in the beginning the network is carefully initialized with random numbers. When this is repeated, the same input set may yield very different networks. Sometimes they differ in performance, one showing good behavior, while others behave badly. Note that it is perfectly possible for a neural network to get stuck in some suboptimal situation; this unfortunate situation can sometimes be avoided through sophisticated techniques.
Neural networks have many parameters, such as learning rate, number of layers, number of neurons, and so on. It always pays off to use different settings for a given problem, thereby trying to find an — at least for the time being — optimal setting in an empiric way. It is not necessarily true that larger networks outperform smaller ones. Not only will the training process take longer, it might

1 Dr W.A. Kosters, kosters@liacs.nl, Leiden University, LIACS, Leiden, The Netherlands

also be the case that the smaller network is more capable of catching the problem at hand.

One difficult problem is to decide whether or not the training has finished. In fact, training can go on as long as one likes, but this sometimes leads to a phenomenon known as 'overfitting': the network gets better and better on the examples it uses during training, but looses its generalizing power. Careful schemata using independent validation sets can avoid these pitfalls. If done carefully, neural networks are perfectly capable of dealing with noisy data, but the danger of overfitting is always present.

Working principles

In this section we describe a simple neural network, officially called a multilayer feed forward neural network. The presentation is kept simple in order to achieve enough understanding of the matter. We focus on what is known as Backpropagation, the most common neural network algorithm. In the next section we shall discuss other 'architectures'.

Network

In our simple network we have one real variable x which is our input. We try to learn $f(x)$, where the function f is unknown to us, but correct pairs $(x,f(x))$ are available. This function (or the value $f(x)$) is called the target. Often x and $f(x)$ are normalized between 0 and 1. For instance, x may be a time variable and $f(x)$ may be the height of the water in a harbor, or x may indicate the distance from the wall and $f(x)$ may be the desired speed of a robot, or x is a zip code and $f(x)$ is the mean income in the corresponding area.

The network has to be trained in such a way that, given input x, it delivers output O with the property that O and $f(x)$ differ as little as possible. The difference $f(x)-O$ between target and output is called the error. Note that the error depends on the particular choice of x.

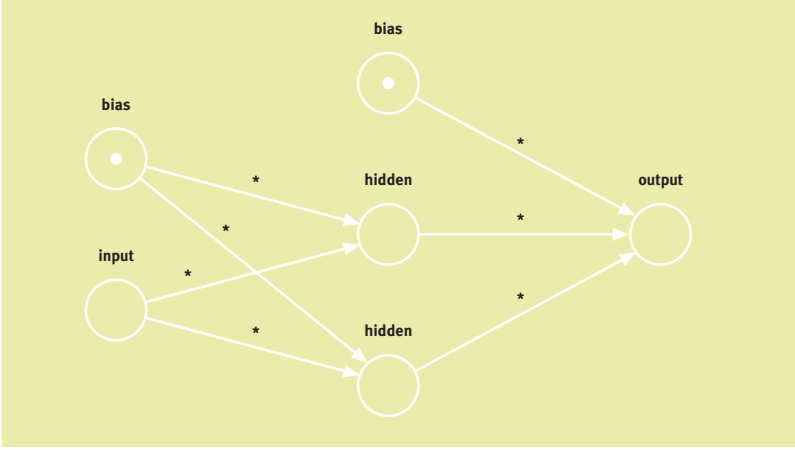


Figure 1
*An example neural network with one input neuron, two hidden neurons and one output neuron. Note the two special bias neurons. Every directed connection has an associated weight, denoted by a *.*

The network itself consists of several so-called neurons. These can be considered as very simple input-output devices. A very simple neural network (see Figure 1) may have one input neuron, one output neuron, and two other neurons: the so-called hidden ones. The input neuron receives the input *x* and hands it to the two hidden neurons. It does so by multiplying *x* by some 'weight', one for every connection. A weight can be viewed as being attached to the directed connection between two neurons. A hidden neuron receives its input, and delivers its output to the output neuron, again multiplying it by some weight. The internal function of a neuron is usually very simple: if its input is low, its output will be low (near zero), and if its input is high, its output will be high (near one). This transfer function is governed by some parameters, one being the 'threshold' or 'bias', which is often implemented by means of extra neurons, one for every layer. Note that in this example network we have four independent weights, or seven if the bias neurons are added. In Figure 1 they are indicated by means of *'s.

Training

The purpose of the training stage is to update the weights in such a way that errors approach, if possible. The change of weights is directed by the errors. Larger errors will lead to larger changes, where the weights that contribute the most are heavily adapted. The so-called learning rate determines the relative change in this process. The algorithm that is used here is called Backpropagation: it propagates the error back through the network, from output to input.

The training usually takes place in the following way. The network is presented with a (random) series of correct input-output pairs. For every pair Backpropagation is used to update (and improve) the weights. After some fixed period, or if the errors are small enough, the training is stopped. The network found can be judged by giving it fresh input-output pairs: the so-called test set. It should come as no surprise that on these pairs the performance of the network will be inferior to that on the training examples. Nevertheless, it gives a good measure of the quality of the network. The performance on a special validation set may be used for the decision to stop training or not: if the error on the validation set starts to increase, this may indicate over fitting.
Many variations are possible. Let us briefly describe some possibilities:

- Training and testing requires a large quantity of examples. In some cases, not enough examples are available. Methods like bootstrapping can be used to artificially increase the number of input-output pairs.
- The behavior of the network is that of a black box. In some cases it is possible to manipulate the internal structure to match the problem at hand. In particular cases this might be extremely difficult, and it is sometimes preferable

to restrict oneself to more complex and hard to interpret networks that give adequate output.

- In line with Occam's razor, which says that in case of several acceptable solutions the simplest one should be preferred, neural network researchers developed all sorts of schemata to decrease network complexity. This results in more complex learning rules, that for instance cause weights to be zero (corresponding to the elimination of weights).
- It might be useful to train several networks at the same time, giving an ensemble of networks. Their independent results can then be combined to obtain a better joined output. In this case statistical techniques can improve the outcome.
- The training can be adapted in many ways. It can or cannot keep track of infeasible solutions. The price paid is that it can become harder to find the proper training algorithm and the proper parameters. But there are many other possibilities. For instance, instead of just using the current input-output pair one can also use information on the previous pair(s) in the form of 'momentum'. Instead of separate incremental training for every example, it is also possible to combine several examples in so-called batch mode. The learning rate can also be adapted during the training process.

In our example we dealt with a situation in which there was only one input variable and one output variable. By adding adequate neurons it is easy to generalize to more complex situations. It is easy to add hidden neurons, which can also be 'layered'. The 'layered' neurons sum all their incoming signals. In this more general setting the error measure is a sum of the squares error. Note that this generalization makes the network much more complex. It is even possible to let the number of neurons grow or shrink during training.
A small problem occurs if one or more variables are not real-valued. As an example, think of a situation where an output variable should contain the day of the week in the form of an integer between 1 and 7. The network may provide 0.314 as output, which can be interpreted as day 3. It is also possible to represent the output by means of seven neurons, each one corresponding to a certain day of the week. Hopefully the network will produce situations where only one of the seven outputs has a value near one, indicating that this day is the proper output.

Data mining with neural networks

In this section we briefly describe some other 'architectures', especially suited to data mining purposes. For more details and yet even more possibilities the interested reader is referred to texts like [Bishop, 1995] and [Silipo, 1999]. These works cover Hopfield nets, Boltzmann machines, specialized networks for time series analysis, and so on.

The Radial Basis Function

The Radial Basis Function (RBF) architecture is especially suited for clustering. Special units try to catch prototypes for each cluster. Variants like probabilistic neural networks are meant for classification tasks. For all these nets a training set of labeled data is necessary. If not, other methods are needed.

Unsupervised learning

In unsupervised learning the network itself tries to find patterns in the data.

Competitive learning

In competitive learning, the neurons compete for an unclassified input example; during the training process their weights gradually converge to some sort of equilibrium, providing an adequate description of the clustering that is hopefully learned.

Self-Organizing Map

One particular successful technique is Kohonen's Self-Organizing Map (SOM, see Section 6.3.2), which can be viewed as a non-linear projection: the number of dimensions of the data diminishes, showing underlying structure. This so-called feature map has the property that inputs which are close together activate neurons that are close together in the network. Other variants pursue Principal Components Analysis (see Section 6.2.4).

Rule extraction

There are several ways to extract rules from neural networks. In a local approach one tries to understand the local behavior of (part of) the network, in a global approach the overall behavior is examined. As an example of the latter, in sensitivity analysis the effect of changes in one variable are studied: how does the output vary?

Conclusion

We may conclude that the family of neural network techniques contains a large number of data mining tools, especially suited for clustering, classification and prediction. See also the CD-rom: tutorial [Andina, 2001]

Literature

- Andina de la Fuente, D. (2001). Artificial Neural Networks. Universidad Politecnica de Madrid, Spain
- Bishop, C.M. (1995). Neural Networks for Pattern Recognition. Oxford University Press
- Silipo, R. (1999). Neural Networks. In: M. Berthold, D.J. Hand. (eds.). Intelligent Data Analysis. Chapter 7. Springer Verlag

6.2.9 Naive Bayes

Arno Siebes[1]

Introduction

One of the important classes of data mining problems is that of classification. The tuples[2] in the database (or better the cases in the real world) belong to a class. The class denotes, e.g. whether a client is a valuable client or not or whether a client is a credit risk or not. In general we can have any finite number of classes.

The goal in classification is, given a set of tuples with their class, to device a classifier that assigns new (unseen) tuples to their correct class. In a business scenario, this might be used to determine which prospective clients will turn out to be valuable clients and which prospective clients will not.

Classification is a problem with a long history both in Statistics and in Computer Science. The result of this is that there is a wide variety of techniques that can be used. For example, discriminant analysis, classification trees, and neural networks. Naive Bayes [Titterington, 1981] is another approach in this arena. Although it is based on an extreme simplification of the problem, it works remarkably well in practice.

Naive Bayes does not deliver an explicit model, it only allows one to assign a class to a (new) tuple. From a data mining point of view, this is not very desirable. However, Naive Bayes is important for two reasons. First and foremost, since it performs so well in practice, it provides a useful yardstick. If a more elaborate (explicit) technique does not perform as well as Naive Bayes, one can doubt the adequacy of the derived model. Secondly, if the explicit model is not that important in an application, Naive Bayes might be a good option.

Naive Bayes

The formal problem in classification is that we have a relation schema $R = \{A_1, \ldots, A_n\}$ with domain $\Re = D_1 \times \ldots \times D_n$ in which D_i is the domain of attribute A–i and a finite set of classes C. Using a finite relation r over R we have to induce a classifier:

$$K: \Re \rightarrow C$$

that assigns the tuples to their correct class.

Since we only have a finite number of examples, i.e. a sample of the whole domain possibly with errors, a probabilistic setting is natural. This means that for a $\vec{t} \in \Re$ we have to estimate $P(c_i \mid \vec{t})$ for all classes c_i *in* C. In fact, assuming

1 Prof Dr A. Siebes, arno@cs.uu.nl, Institute of Information and Computing Sciences, Department of Mathematics and Computer Science, Utrecht University, The Netherlands

2 Sets of variables, rows of the database.

equal misqualification costs (i.e. all misqualifications are equally bad) and 'don't know' answers deemed unacceptable, the (Bayes) optimal solution to the classification problem is:

$argmax_{ci\ in\ C}\ P(c_i \mid \vec{t})$.

In other words, t is assigned to the class with the highest conditional probability. To estimate $P(c_i \mid t)$, note that Bayes theorem gives us:

Formula 1

$$argmax_{ci\ in\ C}\ P(c_i \mid \vec{t}) = argmax_{ci\ in\ C}\ \frac{P\left(\vec{t} \mid c_i\right)P\left(c_i\right)}{P\left(\vec{t}\right)}$$

Formula 2

$$= argmax_{ci\ in\ C}\ P\left(\vec{t} \mid c_i\right)P\left(c_i\right) .$$

For the last equality to hold, we assume that all $\vec{t} \in \Re$ are equally likely; which is not an unreasonable assumption.
Estimating $P(c_i)$ from the data is easy, we can simply count the relative frequency of c_i in r. Estimating $P(\vec{t} \mid c_i)$ is less simple, since $\vec{t}$ might not even occur in r. Therefore, we apply Bayes theorem again. Since $\vec{t} = (t_1, \ldots, t_n)$, Bayes theorem gives us:

$$P\left(t_1, \ldots, t_n \mid c_i\right) = \prod_{j \in \{1,\ldots,n\}} P\left(t_j \mid t_{j+1}, \ldots, t_n, c_i\right)$$

(Where $\prod_j$ stands for the product of the different terms j).

Estimating the $P(t_j \mid t_{j+1}, \ldots, t_n, c_i)$ will in most cases be no easier than estimating the original $P(\vec{t} \mid c_i)$.
To make this estimation simple, we make a (extremely) simplifying assumption: we simply assume that the attributes are independent given the class label. That is, we simply assume that:

$$P(t_j \mid t_{j+1}, \ldots, t_n, c_i) = P(t_j \mid c_i)$$

Estimating the $P(t_j \mid c_i)$ is, of course, easy.
Substituting the assumption in our original specification, we get that we classify tuple t according to:

$$K(t) = argmax_{ci\ in\ C}\ P\left(c_i\right) \prod_{j \in \{1,\ldots,n\}} P\left(t_j \mid c_i\right)$$

This classification rule is naive in that it is only provably correct if the attributes are conditionally independent given the class label. In fact, it is also known as

Idiot's Bayes for this reason. However, in practice it works rather well. Before we go into this, let's look at a small example:

A_1	A_2	A_3	A_4	A_5	Class
1	0	1	0	1	0
0	1	0	1	0	1
1	1	0	0	1	1
0	0	1	1	0	0
1	1	1	0	0	0
0	0	0	1	1	1

How do we classify the tuple (1, 1, 1, 1, 1)? In the table we see that in two of the three rows in which $Class = 0$ we have that $A_1 = 1$. Hence, $P(A_1 = 1 \mid Class = 0)$ is estimated as $2/3$. Similarly, we estimate $P(A_2 = 1 \mid Class = 0) = 1/3$, $P(A_3 = 1 \mid Class = 0) = 1$, $P(A_4 = 1 \mid Class = 0) = 1/3$, and $P(A_5 = 1 \mid Class = 0) = 1/3$. Notice that for $Class = 1$, there are no rows in the table with $A_3 = 1$, hence $P(A_3 = 1 \mid Class = 1) = 0$. So:

Formula 3

$$K(1,1,1,1,1) = argmax_{ci\ in\ C}\ P(c_i) \prod_{j \in \{1,\ldots,n\}} P(t_j|c_i)$$

Formula 4

$$= argmax\{\, 1/2 \times 2/81 \mid Class = 0,\ 1/2 \times 0 \mid Class = 1 \,\}$$

Formula 5

$$= Class = 0$$

Similarly, the tuple (0, 0, 0, 0, 0) gets assigned to class 1.

Why does it work?

Clearly, the assumption that the attributes are conditionally independent, given the class label, will not hold very often in practice. Yet, Naive Bayes works pretty well in practice. How is this possible?

There are two arguments for this phenomenon. The first is based on the observation that we do not have to estimate $P(c_j \mid \vec{t})$ correctly, as long as the correct class gets the largest conditional probability. In other words, errors are not necessarily bad, as long as they don't 'bias' us away from the correct class. The argument is based on the so-called bias/variance decomposition of the error, see [Friedman, 1997].

The second argument is more surprising. In Naive Bayes, we look at the marginal probability distributions instead of the joint distribution. The argument by Garg and Roth [Garg, 2001] is that almost all distributions with the same marginals are very close to the product distribution. This means that the error incurred

using the product distribution instead of the joint distribution is almost always small. In other words, Naive Bayes will almost always do well! See [Garg, 2001] for more details.

References

- Friedman, J.H. (1997). On Bias Variance, 0/1-Loss, and the Curse of Dimensionality. Data Mining and Knowledge Discovery **1** (1):55-78
- Garg, A., D. Roth. (2001). Understanding Probabilistic Classifiers. Proceedings of ECML 2001. Springer Verlag LNCS 2167. pp179-191
- Titterington, D.M., G.D. Murray, L.S. Murray, D.J. Spiegelhalter, A.M. Skene, J.D.F. Habbema, G.J. Gelpke. (1981). Comparison of Discrimination Techniques Applied to a Complex Data Set of Head Injured Patients. Journal of the Royal Statistical Society, Series A **144** (2):144-175

6.2.10 Hidden Markov Models

Roger Boyle[1]

Introduction

This article introduces Hidden Markov Models (HMM), a method especially suited for analysis of sequences of events or time series, recognizing patterns and generating a model.

Frequently, patterns do not appear in isolation but as part of a series in time. This progression can sometimes be used to assist in their recognition. Assumptions are usually made about the time-based process. A common assumption is that the process's state is dependent only on the preceding N states, then we have an order N Markov model. The simplest case is N=1.

Various examples exists where the process states (patterns) are not directly observable, but are indirectly, and probabilistically, observable as another set of patterns. We can then define a HMM. These models have proved to be of great value in many current areas of research, notably speech recognition.

When we can define a HMM, we can tackle three general problems:

- Evaluation: with what probability does a given model generate a given sequence of observations.
- Decoding: what sequence of hidden (underlying) states most probably generated a given sequence of observations.
- Learning: what model most probably underlies a given sample of observation sequences: that is, what are the parameters of such a model?
- We will discuss patterns, hidden patterns, Markov Models and HMMs. The article concludes with three algorithms that can be used in the analysis.

Patterns

Often we are interested in finding patterns which appear over a period of time. These patterns occur in many areas; the pattern of commands someone uses in instructing a computer, sequences of words in sentences, the sequence of phonemes in spoken words — any area where a sequence of events occurs could produce useful patterns.

Consider the simple example of someone trying to deduce the weather from a piece of seaweed. Folklore tells us that 'soggy' seaweed means wet weather, while 'dry' seaweed means sun. If it is in an intermediate state ('damp'), then we cannot be sure. However, the state of the weather is not restricted to the state of the seaweed, so we may say on the basis of an examination that the weather is probably rainy or sunny. A second useful clue would be the state of the weather on the preceding day (or, at least, its probable state) by combining knowledge about what happened yesterday with the observed seaweed state, we might come to a better forecast for today.

1 Dr R. Boyle, roger@comp.leeds.ac.uk, Senior Lecturer in AI, School of Computing, University of Leeds, LS2 9JT, Leeds, United Kingdom, http://www.comp.leeds.ac.uk/

This is typical of the type of system we will consider in this article.

- First we will introduce systems which generate probabilistic patterns in time, such as the weather fluctuating between sunny and rainy.
- We then look at systems where what we wish to predict is not what we observe. The underlying system is hidden. In the above example the observed sequence would be the seaweed and the hidden system would be the actual weather.
- We then look at some problems that can be solved once the system has been modeled. For the above example, we may want to know:
  - What the weather was for a week given each day's seaweed observation.
  - Given a sequence of seaweed observations, is it winter or summer? Intuitively, if the seaweed has been dry for a while it may be summer, if it has been soggy for a while it might be winter.

Generating patterns

Deterministic patterns

Consider a set of traffic lights; the sequence of lights is red - red/amber - green - amber - red. The sequence can be pictured as a state machine, where the different states of the traffic lights follow each other.

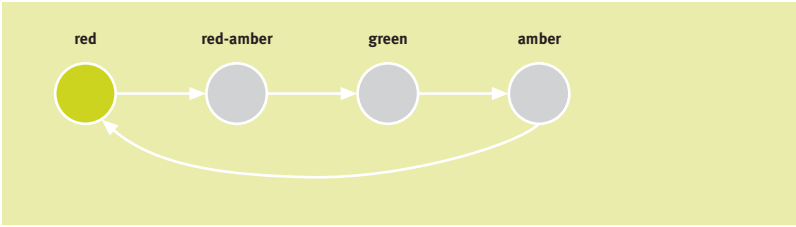


Figure 1
Traffic light sequence.

Notice that each state is dependent solely on the previous state, so if the lights are green, an amber light will always follow: that is, the system is deterministic. Deterministic systems are relatively easy to understand and analyze, once the transitions are fully known.

Non-deterministic patterns

To make the weather example a little more realistic, introduce a third state — cloudy. Unlike the traffic light example, we cannot expect these three weather states to follow each other deterministically, but we might still hope to model the system that generates a weather pattern.

One way to do this is to assume that the state of the model depends only upon the previous states of the model. This is called the Markov assumption and simplifies problems greatly. Obviously, this may be a gross simplification and much important information may be lost because of it.

When considering the weather, the Markov assumption presumes that today's weather can always be predicted solely given the knowledge of the weather of the past few days. Factors such as wind, air pressure, etc. are not considered. In this example, and many others, such assumptions are obviously unrealistic. Nevertheless, since such simplified systems can be subjected to analysis, we often accept the assumption in the knowledge that it may generate information that is not fully accurate.

A Markov process is a process which moves from state to state depending (only) on the previous N states. The process is called an order N model where N is the number of states affecting the choice of the next state. The simplest Markov process is a first order process, where the choice of state is made purely on the basis of the previous state. Notice this is not the same as a deterministic system, since we expect the choice to be made probabilistically, not deterministically. Figure 2 shows all possible first order transitions between the states of the weather example.

Figure 2
States and first order transitions of the weather model.

Notice that for a first order process with M states, there are M^2 transitions between states, since it is possible for any one state to follow another. Associated with each transition is a probability called the state transition probability. This is the probability of moving from one state to another. These M^2 probabilities may be collected together in an obvious way into a state transition matrix. Notice that these probabilities do not vary in time. This is an important (although often unrealistic) assumption.

The state transition matrix below shows possible transition probabilities for the weather example

		weather today		
		sun	*cloud*	*rain*
weather yesterday	*sun*	0.5	0.25	0.25
	cloud	0.375	0.125	0.375
	rain	0.125	0.625	0.375

That is, if it was sunny yesterday, there is a probability of 0.5 that it will be sunny today, and 0.25 that it will be cloudy or rainy. Notice that (because the numbers are probabilities) the sum of the entries for each column is 1.

To initialize such a system, we need to state what the weather was (or probably was) on the day after creation; we define this in a vector of initial probabilities, called the π vector.

$$\begin{pmatrix} \textit{sun} & \textit{cloud} & \textit{rain} \\ 1.0 & 0.0 & 0.0 \end{pmatrix}$$

In this example, we know it was sunny on day 1.

We have now defined a first order Markov process consisting of:

- *states*: three states - sunny, cloudy, rainy.
- *vector*: defining the probability of the system being in each of the states at time o. In a formula:

$$\prod = (\pi_i)$$

the vector of the initial state probabilities.

- *state transition matrix*: the probability of the weather given the previous day's weather. Formula:

$$A = (a_{ij})$$

- with

$$\Pr(x_{j_t} \mid x_{i_{t-1}})$$

Any system that can be described in this manner is a Markov process.

Patterns generated by a hidden process

When a Markov process may not be powerful enough

In some cases the patterns that we wish to find are not described sufficiently by a Markov process. Returning to the weather example, a hermit may perhaps not have access to direct weather observations, but does have a piece of seaweed. Folklore tells us that the state of the seaweed is probabilistically related to the state of the weather. The weather and seaweed states are closely linked. In this case we have two sets of states, the observable states (the state of the seaweed) and the hidden states (the state of the weather). We wish to devise an algorithm for the hermit to forecast weather from the seaweed and the Markov assumption without actually ever seeing the weather.

A more realistic problem is that of recognizing speech; the sound that we hear is the product of the vocal chords, size of throat, position of tongue and several other things. Each of these factors interact to produce the sound of a word, and the sounds that a speech recognition system detects are the changing sound

generated from the internal physical changes in the person speaking.
Some speech recognition devices work by considering the internal speech production to be a sequence of hidden states, and the resulting sound to be a sequence of observable states generated by the speech process that at best approximates the true (hidden) states. In both examples it is important to note that the number of states in the hidden process and the number of observable states may be different. In a three state weather system (sunny, cloudy, rainy) it may be possible to observe four grades of seaweed dampness (dry, dryish, damp, soggy); pure speech may be described by (say) 80 phonemes, while a physical speech system may generate a number of distinguishable sounds that is either more or less than 80.
In such cases the observed sequence of states is probabilistically related to the hidden process. We model such processes using a hidden Markov model where there is an underlying hidden Markov process changing over time, and a set of observable states which are related somehow to the hidden states.

Hidden Markov Models

The diagram below shows the hidden and observable states in the weather example. It is assumed that the hidden states (the true weather) are modeled by a simple first order Markov process, and so they are all connected to each other.

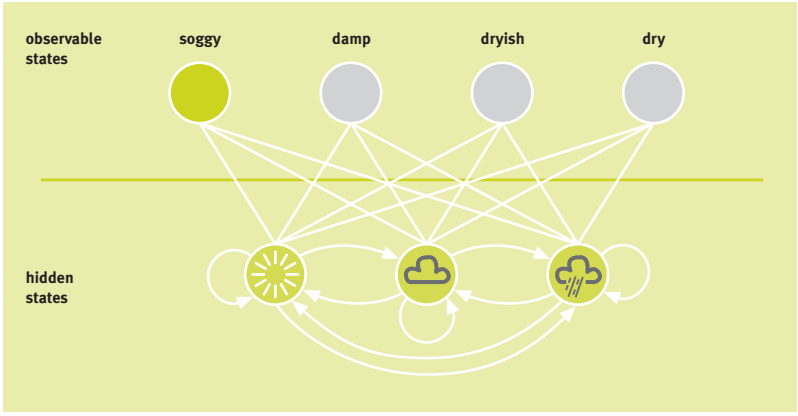


Figure 3
Hidden and observable states.

The connections between the hidden states and the observable states represent the probability of generating a particular observed state given that the Markov process is in a particular hidden state. It should thus be clear that all probabilities 'entering' an observable state will sum to 1, since in the above case it would be the sum of Pr(Obs|Sun), Pr(Obs|Cloud) and Pr(Obs|Rain).
In addition to the probabilities defining the Markov process, we therefore have another matrix, termed the confusion matrix, which contains the probabilities of the observable states given a particular hidden state. For the weather example the confusion matrix might be

$$
\begin{array}{cc} & \begin{array}{ccc} \multicolumn{3}{c}{\textit{weather}} \\ \textit{sun} & \textit{cloud} & \textit{rain} \end{array} \\ \begin{array}{c} \\ \textit{seaweed} \end{array} & \begin{array}{c} \textit{dry} \\ \textit{dryish} \\ \textit{damp} \\ \textit{wet} \end{array} \begin{pmatrix} 0.60 & 0.25 & 0.05 \\ 0.20 & 0.25 & 0.10 \\ 0.15 & 0.25 & 0.35 \\ 0.05 & 0.25 & 0.50 \end{pmatrix} \end{array}
$$

Notice that the sum of each matrix row is 1.
The formula for the confusion matrix being:

$B = (b_{ij})$

with probabilities

$\Pr(y_i \mid x_i)$

Thus, a HMM is a triple (π,A,B). Besides the initial state vector π and the state transition matrix A, a confusion matrix B is described.
Each probability in the state transition matrix and in the confusion matrix is time independent: that is, the matrices do not change in time as the system evolves. In practice, this is one of the most unrealistic assumptions of Markov models about real processes.

Usage associated with HMMs

Once a system can be described as a HMM, three problems can be solved. The first two are pattern recognition problems: evaluation and decoding. The third problem is a learning task.

Evaluation: finding the probability of an observed sequence

Consider the problem where we have a number of HMMs (a set of (μ,A,B) triples) describing different systems, and a sequence of observations. We may want to know which HMM most probably generated the given sequence. For example, we may have a 'Summer' model and a 'Winter' model for the seaweed, since behaviour is likely to be different from season to season. We may then hope to determine the season on the basis of a sequence of dampness observations.
To calculate the probability of an observation sequence given a particular HMM, and hence choose the most probable HMM, the forward algorithm is used.
This type of problem occurs in speech recognition where a large number of Markov models will be used, each one modeling a particular word. An observation sequence is formed from a spoken word, and this word is recognized by identifying the most probable HMM for the observations.

Decoding: finding the most probable sequence of hidden states given some observations

Another related problem, and the one usually of most interest, is to find the hidden states that generated the observed output. In many cases we are interested in the hidden states of the model, since they represent something of value that is not directly observable.
Consider the example of the seaweed and the weather; a blind hermit can only sense the seaweed state, but needs to know the weather, i.e. the hidden states. To determine the most probable sequence of hidden states given a sequence of observations and a HMM, the Viterbi algorithm is used.
A widespread application of the Viterbi algorithm is in Natural Language Processing, to tag words with their syntactic class (noun, verb, etc.) The words in a sentence are the observable states and the syntactic classes are the hidden states (note that many words, such as wind, fish, may have more than one syntactical interpretation). By finding the most probable hidden states for a sentence of words, we have found the most probable syntactic class for a word, given the surrounding context. Thereafter we may use the primitive grammar so extracted for a number of purposes, such as recapturing 'meaning'.

Learning: generating a HMM from a sequence of observations

The third, and much the hardest, problem associated with HMMs is to take a sequence of observations (from a known set), known to represent a set of hidden states, and fit the most probable HMM; that is, determine the (μ,A,B) triple that most probably describes what is seen. Here the forward-backward algorithm is used. The forward-backward algorithm is of use when the matrices A and B are not directly (empirically) measurable, as is very often the case in real applications.

Algorithms for HMM's

We will only briefly discuss the algorithms involved, for a more comprehensive discussion, including examples we refer to the complete HMM course that is included on the CD-rom.

Forward algorithm definition

We use the forward algorithm to find the probability of an observed sequence given a HMM. It exploits recursion in the calculations to avoid the necessity for exhaustive calculation of all paths through the execution trellis.
Given this algorithm, it is straightforward to determine which of a number of HMMs best describes a given observation sequence: the forward algorithm is evaluated for each, and that giving the highest probability selected.

The forward algorithm is used to calculate the probability of a T long observation sequence

$$Y^{(k)} = y_{k1}, \ldots, y_{kT}$$

where each of the y is one of the observable set. Intermediate probabilities (α's) are calculated recursively by first calculating α for all states at t=1.

$$\alpha_1(j) = \pi(j).b_{k1j}$$

Then for each time step, t = 2, ..., T, the partial probability α is calculated for each state

$$\alpha_{t+1}(j) = \sum_{i=1}^{n} (\alpha_t(i)\, a_{ij})\, b_{k_t j}$$

that is, the product of the appropriate observation probability and the sum over all possible routes to that state, exploiting recursion by knowing these values already for the previous time step.
Finally, the sum of all partial probabilities gives the probability of the observation, given the HMM, λ.

$$\Pr\left(Y^{(k)}\right) = \sum_{j=1}^{n} \alpha_T(j)$$

To recap, each partial probability (at time t > 2) is calculated from all the previous states. The recursion reduces the computational complexity.

Viterbi algorithm definition

The Viterbi algorithm provides a computationally efficient way of analyzing observations of HMMs to recapture the most likely underlying state sequence. It exploits recursion to reduce computational load, and uses the context of the entire sequence to make judgments, thereby allowing good analysis of noise. In use, the algorithm proceeds through an execution lattice calculating a partial probability for each cell, together with a back-pointer indicating how that cell could most probably be reached. On completion, the most likely final state is taken as correct, and the path to it traced back to t=1 via the back pointers.

The algorithm may be summarized formally as:

For each i,, i = 1, ... , n, let:

$$\delta_1(i) = \pi(i)\, b_{k_1 i}$$

This initializes the probability calculations by taking the product of the initial hidden state probabilities with the associated observation probabilities.

For t = 2, ..., T, and i = 1, ... , n let:

$$\delta_t(i) = \max_j(\delta_{t-1}(j)\, a_{ji} b_{k_t\, i})$$

$$\phi_t(i) = \text{argmax}_j(\delta_{t-1}(j) a_{ji})$$

Thus, determining the most probable route to the next state, and remembering how to get there. This is done by considering all products of transition probabilities with the maximal probabilities already derived for the preceding step. The largest such is remembered, together with what provoked it.
Let:

$$i_t = \text{argmax}(\delta_T(i))$$

Thus, determining which state at system completion (t=T) is the most probable.
For t = T - 1, ..., 1
Let:

$$i_t = \phi_{t+1}(i_{t+1})$$

Thus, backtracking through the lattice, following the most probable route. On completion, the sequence i_1 ... i_T will hold the most probable sequence of hidden states for the observation sequence in hand.

As with the forward algorithm, there is a reduction in computational complexity by using recursion.
The Viterbi algorithm has the very useful property of providing the best interpretation given the entire context of the observations.
In other words, the Viterbi algorithm will look at the whole sequence before deciding on the most likely final state, and then 'backtracking' through the calculations to indicate how it might have arisen. This is very useful in 'reading through' isolated noise garbles, which are very common in live data.

Forward-backward algorithm

The 'useful' problems associated with HMMs are those of evaluation and decoding. They permit either a measurement of a model's relative applicability, or an estimate of what the underlying model is doing (what 'really happened').
It can be seen that they both depend upon foreknowledge of the HMM parameters: the state transition matrix, the observation matrix, and the π vector.
There are, however, many circumstances in practical problems where these are not directly measurable, and have to be estimated. This is the learning problem.

The forward-backward algorithm permits this estimate to be made on the basis of a sequence of observations known to come from a given set that represents a known hidden set following a Markov model.
An example may be a large speech processing database, where the underlying speech may be modeled by a Markov process based on known phonemes, and the observations may be modeled as recognizable states (perhaps via some vector quantization), but there will be no (straightforward) way of deriving empirically the HMM parameters.
The forward-backward algorithm is not unduly hard to comprehend, but is more complex in nature than the forward algorithm and the Viterbi algorithm. For this reason, it will not be presented here in full (any standard reference on HMMs will provide this information (see the References).
In summary, the algorithm proceeds by making an initial guess of the parameters (which may well be entirely wrong) and then refining it by assessing its worth, and attempting to reduce the errors it provokes when fitted to the given data. In this sense, it is performing a form of gradient descent, looking for a minimum of an error measure.
It derives its name from the fact that, for each state in an execution trellis, it computes the 'forward' probability of arriving at that state (given the current model approximation) and the 'backward' probability of generating the final state of the model, again given the current approximation. Both of these may be computed advantageously by exploiting recursion, much as we have seen already. Adjustments may be made to the approximated HMM parameters to improve these intermediate probabilities, and these adjustments form the basis of the algorithm iterations.

Conclusion

HMMs have proved to be of great value in analyzing real systems, and are used for evaluation, decoding and learning tasks. Their usual drawback is the oversimplification associated with the Markov assumption that a state is only dependent on predecessors, and that this dependence is time independent (constant).

References

– Boyle, R., S. Hanlon. (2001). The University of Leeds' Course on Hidden Markov Models. University of Leeds, United Kingdom, http://www.scs.leeds.ac.uk/scs-only/teaching-materials/HiddenMarkovModels/html_dev/main.html

A full exposition on HMMs may be found in:

– Rabiner, L.R., B.H. Juang. An Introduction to HMMs. iEEE ASSP Magazine **3**:4-16. This document is an excerpt from The University of Leeds' course on Hidden Markov Models, also included on the CD-rom: HMM course

6.2.11 Belief networks/Bayesian networks

Wim Wiegerinck[1], *Tom Heskes*[2]

Introduction

In modeling real world tasks, one inevitably has to deal with uncertainty. This uncertainty is due to the fact that many facts are unknown and or simply ignored and summarized. Suppose that one morning you find out that your grass is wet. Is it due to rain, or is it due to the sprinkler? If there is no other information, you can only talk in terms of probabilities. In a probabilistic model approach, you could try to enumerate the states of all variables (grass: wet or dry; rained: true or false; sprinkler: on or off), and assign probabilities to each combination of states. Ideally, these probabilities will be proportional to the relative frequencies of the occurrence of the combinations of states.

The elegance of the probabilistic approach resides in the fact that the probabilistic model on these three variables is correct, consistent and automatically includes context dependency. For instance, you can use the model to compute the probability that it has rained in the context that the grass is wet. You will find an increase in the probability that it has rained. However, in the context that the grass is wet *and* that the sprinkler has been left on, the probability that it has rained is generally (for sensible choices of the conditional probabilities) lower.

In systems that are rule-based rather than based on probability theory context dependency is not fully modeled. In such systems invalid conclusions can be drawn easily. For instance, in a system with context free rules, concatenation of the rule: 'sprinkler on' implies 'wet grass' with the rule: 'wet grass' implies 'it has rained', will lead to the incorrect conclusion that 'sprinkler on' implies 'it has rained'.

A drawback of probabilistic models is their computational complexity. In problems with many variables the approach in which all combinations of states are enumerated in the model will lead to huge computational problems. The reason is that the number of combinations of states grows exponentially with the number of variables. Even if one manages to parameterize the probabilities in an efficient way, the problem is still not easily solved: inference (i.e. computing probabilities of variables of interest) requires the summation over all (exponentially many) states of the remaining variables.

Graphical models provide a remedy. They include Bayesian networks (also known as belief networks), Hidden Markov models, Markov fields, naive Bayes and many others. In graphical models, the probability distribution is defined in terms of local quantities, involving only a few variables. In particular, the local structure can be represented by a graph; hence the name graphical model. The local quantities are glued together according to the laws of probability theory, such that they define a unique and consistent global probability distribution.

1 Dr T. Heskes, tom@mbfys.kun.nl, SNN, Nijmegen University, Nijmegen, The Netherlands

2 Dr W. Wiegerink, wimw@mbfys.kun.nl, SNN, Nijmegen University, Nijmegen, The Netherlands
http://www.snn.kun.nl/

In graphical models the simplifying assumption is that variables that are not directly connected in the graph are (conditionally) independent. Although this may seem a severe restriction, it is in many cases a quite natural and intuitive one. Suppose that we want to extend the wet grass model with a variable representing the neighbor's grass. Now in general, the states of the neighbor's grass and your own grass are dependent. The reason is simply that, if your grass is wet, it probably has rained, and thus your neighbor's grass will also be wet. Natural assumptions, however, are that given the state of 'rained', the states of both grasses are independent, and in the same context that the state of your neighbors grass is independent of the state of your sprinkler. Note that if you do not know the state of 'rained', the fact that your sprinkler has been on may imply a reduced probability of 'rained', thus a reduced probability of your neighbor's grass being wet. Again, this is an example of the context dependency of probabilistic models, referred to as 'explaining away' (see also the example below).
The (conditional) independencies in graphical models do not only simplify the representation of these models, they are also exploited by efficient inference algorithms. This facilitates the practical usage of graphical models. It should be stressed that these inference algorithms are exact according to the laws of probability theory. In large, complex networks, however, even these algorithms may become computationally too demanding. In such cases, approximate inference methods such as stochastic sampling are needed.
A Bayesian network is a particular type of graphical model, frequently used in applications of artificial intelligence for building probabilistic expert systems. An appealing feature of Bayesian networks is that their graphical structure can often be loosely interpreted as the result of direct causal relations between variables. In domains with lots of causal relations, such as medical diagnosis (diseases cause symptoms), human experts are usually able to express their domain knowledge in the graphical structure of the network. The parameters of the network are the conditional probabilities of effects given the state of their direct causes. Bayes' rule

$$P\left(cause\middle|effect\right)=\frac{P\left(effect\middle|cause\right)P\left(cause\right)}{P\left(effect\right)} \quad ,$$

arises when we want to reason from symptom (effect) to disease (cause) and thus have to 'invert' the probabilities.
The exact values of the conditional probabilities are often more difficult for human experts to determine. Fortunately, these conditional probabilities are relatively easy to estimate from data: in its simplest version learning in Bayesian networks (given their graphical structure) amounts to straightforward frequency counting. If the experts are not able to specify the graphical structure, there is the real challenge of learning the structure of Bayesian networks from data.

Bayesian networks and probability theory

The mathematics of Bayesian networks is most easily explained through an example. So let us consider the wet-grass example with four variables *R* (Rained), *S* (Sprinkler being on), *G* (Grass wet) and *N* (Neighbors grass wet). Each variable can be in two states, true or false. The joint probability distribution $P(R,S,G,N)$ is a table with 16 entries. The table is normalized, i.e.

$$\sum_{R=\{t,f\},\ldots,N=\{t,f\}} P(R,S,G,N)=1$$

With this table we can compute any variable of interest, e.g. $P(R=f \mid N=t)$, the probability that it has not rained given that my neighbor's grass is wet, reads

$$P(R=f \mid N=t)=\frac{P(R=f,N=t)}{P(N=t)}=\frac{\sum_{S=\{t,f\},G=\{t,f\}} P(R=f,S,G,N=t)}{\sum_{R=\{t,f\},S=\{t,f\},G=\{t,f\}} P(R,S,G,N=t)}$$

Now according to probability theory, we can write any joint probability distribution as a product of conditional distributions:

$$P(R,S,G,N)=P(R)P(S \mid R)P(G \mid R,S)P(N \mid R,S,G)\ .$$

In a Bayesian network, conditional independencies are assumed. For instance, we may assume for our model that being given (only) the state of Rained does not influence probability of the Sprinkler being on (an independency not yet discussed in the introductory section),

$$P(S \mid R)=P(S)$$

and given the state of Rained, the additional knowledge of the state of sprinkler and or the state of your grass will not influence the probability of the neighbor's grass being wet,

$$P(N \mid R,S,G)=P(N \mid R)$$

With these model assumptions (which may be completely wrong, but that is another issue), the joint probability in our Bayesian network is of the form

Formula 1

$$P(R,S,G,N)=P(R)P(S)P(G \mid R,S)P(N \mid R)$$

In Figure 1, the corresponding (directed acyclic) graph or DAG is visualized. Each variable is represented as a node. Each combination of a node and its incoming

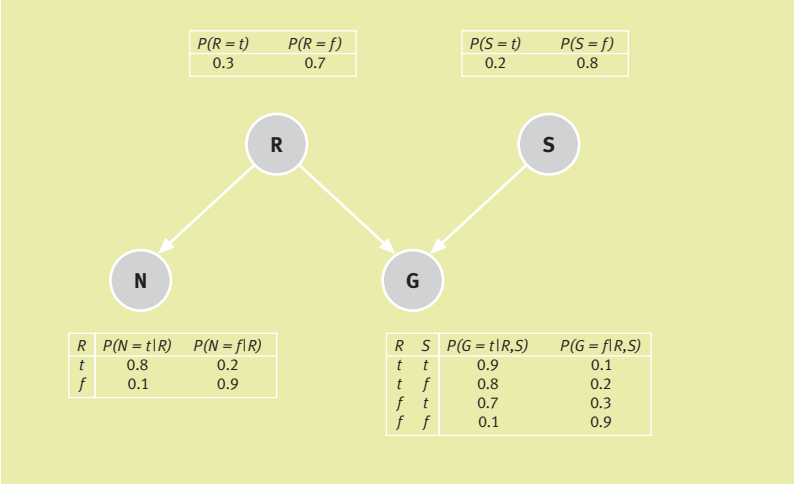


Figure 1
Graphical representation and conditional probability tables for the 'wet grass' model.

arrows corresponds to a conditional probability distribution. For example, the node *G* with incoming arrows from *R* and *S* corresponds to the distribution $P(G|R,S)$. It can be shown that there is a 1-1 correspondence between DAGs and factorizations of the joint distribution into conditionals. If there is no conditional independence assumption, the graph is fully connected. Independency assumptions lead to deletion of arrows.

Furthermore, it is clear that if we define the conditional probabilities on the right hand side of equation (1), the joint probability is fully specified. The conditional probability distributions that we need to complete the definition of our model are given in Figure 1.

Our Bayesian network is now ready for carrying out inference. For instance, we can compute the two marginal probabilities of the grasses *G* and *N* being wet:

$$P(G=t) = \sum_{R=\{t,f\},S=\{t,f\}} P(R)P(S)P(G=t|R=t,S) = 0.4$$

$$P(N=t) = \sum_{R=\{t,f\}} P(R)P(N=t|R) = 0.31$$

If we know that it has rained we can compute how these probabilities change through conditioning:

$$P(G=t|R=t) = \sum_{S=\{t,f\}} P(S)P(G=t|R=t,S) = 0.82$$

$$P(N=t|R=t) = \frac{P(S=t,G=t)}{P(G=t)} = 0.8$$

So, given that it has rained, the probabilities of both grasses being wet increase, as they should. On the other hand, if it is found that your grass is wet,

we can compute the probability that it has rained, as well as the probability that the sprinkler has been on.

$$P(R=t \mid G=t)=\frac{P(R=t,G=t)}{P(G=t)}=0.615\ ,$$

$$P(S=t \mid G=t)=\frac{P(S=t,G=t)}{P(G=t)}=0.38\ .$$

So, both probabilities are higher than their a priori probabilities $P(R=t) = 0.3$ and $P(S=t) = 0.2$, respectively. In other words, the probabilities of both causes R and S increase, if their effect G is found to be true.
If it is observed now that the neighbor's grass is also wet, we obtain the probabilities

$$P(R=t \mid G=t,N=t)=\frac{P(R=t,G=t,N=t)}{P(G=t,N=t)}=0.9274\ ,$$

$$P(S=t \mid G=t,N=t)=\frac{P(S=t,G=t,N=t)}{P(G=t,N=t)}=0.2498\ .$$

Clearly, the possible cause R becomes more likely since additional evidence $N=t$ is found. Since the increased probability of R provides an explanation of $G=t$, we no longer need the explanation provided by the other cause, which therefore receives a lower probability. It is said that S is 'explained away'.

Bayesian networks and causal modeling?

A very important modeling issue in Bayesian networks is the direction of the arrows. It is instructive to consider what the consequences in the model are, if we reverse the incoming arrows to G. One could say: (1) wet grass indicates rain, therefore there should be an arrow from G to R and (2) wet grass indicates that the sprinkler has been on, therefore there should be an arrow from G to R. This model would lead to similar insensible conclusions as the concatenation of deterministic rules in the introductory section: if it has rained, the grass is probably wet, and this leads to an increased probability of the sprinkler being on.
In a similar way, it is instructive to consider the consequences of reversing the outgoing arrows from R, i.e. such that arrows point from N to R and from G to R. Given that it has rained, the probabilities of N and G both increase. Now suppose that you next observe that the neighbor's grass is wet. Then according to the model, the neighbor's wet grass explains the rain, and therefore the probability of your grass being wet decreases! This does definitely not correspond to common sense and therefore we deem the original model, with arrows from cause to effect, more sensible than the one with reversed arrows.
As we have seen, in general it is a good rule of thumb to construct a Bayesian

network from cause to effect. You start with nodes that represent independent root causes, then model the nodes they influence, and so on until you end at the leaves, i.e. the nodes that have no direct influence on other nodes. For this procedure, it is often useful to have a 'story' in mind.
Sometimes this procedure fails, because it is too difficult to tell what is cause and what is effect. Is someone's behavior a result of his environment, or is the environment a reaction to his behavior? In such a case, you should just avoid the philosophical dispute, and return to the basics of Bayesian networks: a Bayesian network is not a model for causal relations, but a joint probability model. The structure of the network represents the conditional independence assumptions in the model and nothing else.
In practice it is often difficult to decide whether two nodes are really (conditionally) independent. Usually, this is a matter of simplifying model assumptions. In the case of the wet grass model, one could easily argue that *N* and *G* are still dependent, even if we know that it has not rained, e.g. due to humidity or other weather conditions that increase the probability of both grasses being wet simultaneously. In the true world, all nodes should be connected. In practice, reasonable (approximate) assumptions are needed to make the model simple enough to handle, but still powerful enough for practical usage.

Software and further reading

SNN Nijmegen has developed a freeware tool for building and computing with Bayesian networks, called BayesBuilder, available for both Windows and Linux platforms, which is included on the CD. Its most recent version can be downloaded from the site http://www.snn.kun.nl/nijmegen/bayesbuilder.html
The reference work on Bayesian networks is the book by [Pearl,1988]. [Cowell, 1999] is more up to date and somewhat less technical. The tutorial by [Heckerman, 1998] specifically treats learning in Bayesian networks. The collection 'Learning in Graphical Models' as a whole gives an impression of the state of the art and current challenges. Lots of information can be found on the internet, with http://www.auai.org, the homepage of the Association for Uncertainty in Artificial Intelligence, as a good starting point. See also Section 2.3.2, Decision support for medical diagnosis.

References

- Cowell, R., A. Dawid, S. Lauritzen, D. Speigelhalter. (1999). Probabilistic Networks and Expert Systems. Springer Verlag, Berlin
- Heckerman, D.(1998). A Tutorial on Learning with Bayesian Networks. In: M. Jordan (ed.). Learning in Graphical Models **89** of NATO ASI, series D. Behavioural and Social Sciences:301-354. Kluwer
- Pearl, J. (1988). Probabilistic Reasoning in Intelligent systems: Networks of Plausible Inference. Morgan Kaufmann, San Francisco, CA

6.2.12 Association rules

Arno Siebes[1]

Introduction

One of the important categories of data mining problems is that of associations between attributes. This gives useful insight into such diverse business problems as product cross-selling, web site perception, and decision problems.
There are two ways to look at attribute associations. The first is on the attribute level, i.e. one looks for statistical dependencies between the attributes. Bayesian networks can be used for such associations. The second way is at the value level. Association rules [Agrawal, 1993] are the premier tool for this class of problems.
Informally, association rules can be seen as if-then rules: if a person buys a newspaper, he or she also buys chocolate. The twist association rules bring to classical if-then rules is a conditional probability. If a person buys a newspaper, there is a probability that he or she will also buy chocolate. This conditional probability is known as confidence in the literature.
Another measure that is generally associated with association rules is that of support: the fraction of customers for whom the rule holds, or rather the relative number of customers buying all items occurring in the rule (the so-called underlying itemset). If there is just one customer that buys newspapers and he or she also happens to buy chocolate, the association rule is not very interesting.
Next to being a measure of interestingness, the support of a rule also plays a key role in the standard algorithms for association rule discovery. Given thresholds minsup and minconf for the support and confidence, these algorithms compute all association rules, whose support and confidence exceed these thresholds. For these rules the underlying itemset is called frequent.
Originally, association rules were introduced in a binary, non-temporal setting. For example, one considered a collection of transactions at the check-out of a store only recording whether a certain item was bought or not. Later, many extensions were introduced, e.g. sequences (time), hierarchical clusters of items, and item-counts.

Association rules: the binary case

The traditional setting for association rules [Agrawal, 1994] consists of a set of transactions in which each transaction is a set of items. Translated to a relational setting, we have a table r with schema $R = \{ A_1, \ldots A_n \}$, in which each A_i is a binary attribute. The attributes correspond to the items, the rows in the table to the transactions; a row has value 1 for an attribute if and only if the transaction contains that item.

1 Prof Dr A. Siebes, arno@cs.uu.nl, Institute of Information and Computing Sciences, Department of Mathematics and Computer Science, Utrecht University, The Netherlands

For $X, Y \subseteq R$, with $X \cap Y = \varnothing$ let:

- $s(X)$ denote the support of X, i.e. the number of tuples that have value 1 for all attributes in X
- for an association rule $X \rightarrow Y$, define:
  - the support is $s(XY)$
  - the confidence is $s(XY)/s(X)$

The problem is to find all association rules that match or exceed the user defined lower thresholds for confidence, minconf, and support minsup. There are two thresholds we have to satisfy, so the basic algorithm is [Agrawal, 1994]:

1 Find all sets Z whose support exceeds the minimal threshold. These sets are called frequent (or large) sets.
2 Then test for all non-empty subsets X of frequent sets Z whether the rule $X \rightarrow Z \setminus X$ $(=Y)$ holds with sufficient confidence.

The A Priori algorithm

The first problem is then: how do we find the frequent sets? In principle, we have to check all subsets of R. However, this is not possible, since checking only 100 items, we would need (roughly) 4×10^{18} years, which (far) exceeds the age of the universe!

Luckily, we have the following observation [Agrawal, 1994]:

Z can only be frequent, if all its (non-empty) subsets are frequent!

This is known as the A Priori property. In other words, we can search level wise for the frequent sets. The level is the number of items in the set. Denote by $C(i)$ the sets of i items that are potentially frequent (the candidate sets) and by $F(i)$ the frequent sets of i items.

To avoid multiple generations of the set XY (which means unnecessary work), one places an order on the items and uses the induced lexicographical order on sets of items in the generation of candidate sets.

Let's look at an example. Minsup equals 2 and the data is given by:

Tuple id	I_1	I_2	I_3	I_4	I_5
1	1	1	0	0	1
2	0	1	0	1	0
3	0	1	1	0	0
4	1	1	0	1	0
5	1	0	1	0	0
6	0	1	1	0	0
7	1	0	1	0	0
8	1	1	1	0	1
9	1	1	1	0	0

On the first level we find:

Itemset	Support	Check
$\{I1\}$	6	Ok
$\{I2\}$	7	Ok
$\{I3\}$	6	Ok
$\{I4\}$	2	Ok
$\{I5\}$	2	Ok

On the second level we get:

Itemset	Support	Check
$\{I1, I2\}$	4	Ok
$\{I1, I3\}$	4	Ok
$\{I1, I4\}$	1	No
$\{I1, I5\}$	2	Ok
$\{I2, I3\}$	4	Ok
$\{I2, I4\}$	2	Ok
$\{I2, I5\}$	2	Ok
$\{I3, I4\}$	0	No
$\{I3, I5\}$	1	No
$\{I4, I5\}$	0	No

The third level yields:

Itemset	Support	Check
$\{I1, I2, I3\}$	2	Ok
$\{I1, I2, I5\}$	2	Ok

Clearly there are no sets that qualify for the fourth level and we are done.

If we assume that r is sparse (by far the most values are 0), then we expect that the frequent sets have a maximal size k with $k << |R|$. If that expectation is met, we have a worst case complexity far below that of the naïve algorithm.

Generating association rules from the frequent sets is done as follows [Agrawal, 1994].

Generate association rules
For each frequent set X **do**
For all non-empty $Y \subset X$ **do**
If $s(X) \,/\, s(X \setminus Y) \geq$ minconf **then**
Output $X \setminus Y \rightarrow Y$

Continuing our example, we get: One of the frequent sets is $\{ I_1, I_2, I_5 \}$.
This generates:

Itemset	Rule	Confidence
$\{I1, I2\}$	$\{I1, I2\} \rightarrow I5$	$2/4 = 50\%$
$\{I1, I5\}$	$\{I1, I5\} \rightarrow I2$	$2/2 = 100\%$
$\{I2, I5\}$	$\{I2, I5\} \rightarrow I1$	$2/2 = 100\%$
$\{I1\}$	$I1 \rightarrow \{I2, I5\}$	$2/6 = 33\%$
$\{I2\}$	$I2 \rightarrow \{I1, I5\}$	$2/7 = 29\%$
$\{I5\}$	$I5 \rightarrow \{I1, I2\}$	$2/2 = 100\%$

Clearly, this algorithm is again exponential. For every X, we consider all $2^{|X|} - 1$ non-empty subsets Y of X. However, as long as $|X| \leq k << |R|$, this is not necessarily a problem.
Quite often one generates only those association rules with a singleton Y, which turns the generation algorithm linear.

Complexity and the databases

We only looked at the complexity with regard to the number of attributes (items) of our table. However, this is not the only aspect: what about the role of the database?

- If we check each itemset separately, we need as many passes over the database as there are candidate frequent sets.
- If at each level we first generate all candidates and check all of them in one pass, we need as many passes as the size of the largest candidate set.

If the database does not fit in main memory, such passes are costly in terms of input/output time. Can we do it more cheaply?
The only way to reduce the costs is to reduce the number of passes, since we can't build the candidates for multiple levels in one go (this would lead to an exponential number of candidates).
We will discuss two related options:

- partition the database into portions p_i that do fit in main memory; recall that $\cup p_i = db$ and $p_i \cap p_j = \emptyset$ [Savarese, 1995].
- use sampling [Toivonen, 1996].

Using partitions

Assume (without loss of generality) that the minimum support is given as a percentage of tuples. The idea is then as follows: x
Generate a partitioning P of the database
For each $p \in P$ **do**
read p into main memory
compute all frequent item sets on p
in the sets $F_p(i)$
For all levels i **do**
$C(i) = \cup_{p \in P} F_p(i)$
check all candidates in $C = \cup_i C(i)$ in one pass.

We can prove that all candidates generated on the complete database will also be generated by the partitioned approach. However, there may be candidates generated by the partitioned approach that would not be generated on the complete database. This happens when either one or more of the partitions are a biased sample or when one or more of the partitions are too small. In other words, ideally the partitions should be good random samples. This observation begs the question: why don't we use one random sample?

Mining from one sample

If we mine just one sample for association rules, we will make mistakes:

- We will find rules that do not hold on the complete data set.
- We will not find rules that do hold on the complete data set.

Clearly, the probability of such errors depend on the size of the sample. Can we say something about this probability and its relation to the size?

If we want the probability that the estimated (relative) cover of an itemset is more than ε from the true value to be smaller than δ we can calculate ([Toivonen, 1996]) the following sample size using Chernoff bounds[2]:

$$|Sa| \geq \frac{1}{2\varepsilon^2} ln \frac{2}{\delta} \quad .$$

To get a feeling for the required sample sizes, consider the following table:

2 Chernoff bounds constrain the total probability of a random variable situated far from the mean.

ε	δ	$\lvert Sa \rvert$
0.01	0.01	27000
0.01	0.001	38000
0.01	0.0001	50000
0.001	0.01	2700000
0.001	0.001	3800000
0.001	0.0001	5000000

If we want to have a low probability (say, μ) of missing association rules on the sample, we can mine with a lower threshold t'. How much lower should we set it for a given sample? We can calculate a new threshold [Toivonen, 1996]:

$$t' = s_{db}(Z) - \sqrt{\frac{1}{2|Sa|} \ln \frac{1}{\mu}} \quad .$$

In other words, we should lower the threshold by $\sqrt{\frac{1}{2|Sa|} \ln \frac{1}{\mu}}$.

The main idea of using just one sample to mine for association rules is now as follows:

- Draw (with replacement) a sample of sufficient size.
- Compute the set F of all frequent sets on this sample, using the lowered threshold.
- Check the support of the elements of F on the complete database.

This means that we have to scan the complete database only once. Although, taking the random sample may require a complete database scan also!
There is the possibility that we will miss frequent sets. Can we check whether we are missing results in the same database scan? If we check not only F for frequency, but $F \cup Bd(F)$ and warn when an element of $Bd(F)$ turns out to be frequent, we know that we might have missed frequent sets. In fact, if we define

$$Cl(F) = F \cup Bd(F) \cup Bd(F \cup Bd(F)) \cup \ldots$$

we know that all frequent sets have to lie in $Cl(F)$.

Drowning in association rules

In practice, association rules suffer from an embarrassment of richness: one often gets too many results. The number of association rules one discovers is inversely related to both minsup and minconf. If one sets these thresholds (too) high, one only discovers already well-known associations. If one lowers the thresholds, the number of discovered rules grows dramatically. Getting more

results than tuples in the database is not unheard of! Recall our earlier example in which the frequent itemset $\{I_1, I_2, I_5\}$ generated 6 association rules.
If all discovered rules were interesting, the fact that one gets an overload would be an(perhaps unfortunate) fact of life. However, many of the rules convey little or no useful information. Suppose you discover that 60% of the people that buy bread also buy cheese. How interesting is this, if you know that 60% of all people buy cheese?
There are two approaches to this flood of results, i.e. pre-computing and post-processing. Pre-computing means that we try to generate less rules. Post-processing means that the resulting set of rules is filtered or ordered in such a way that the user only has to consider the more interesting rules. Historically post-processing was considered first, therefore we start with this approach.

Post-processing the result set

If there are so many rules, it may be a good idea to order the results. A simple idea is:

- Order the rules by consequent.
- Per consequent:
  - order the rules on the length of the antecedent;
  - present sequences of rules, $\{A_1 \rightarrow C, \{A_1, A_2\} \rightarrow C, \ldots\}$, while replicating rules when necessary.

We can also use confidence and support, they define a partial order on the set of rules:

- $r_1 \leq_{sc} r_2$ iff
  - $s(r_1) < s(r_2) \wedge conf(r_1) = conf(r_2)$ or
  - $s(r_1) = s(r_2) \wedge conf(r_1) < conf(r_2)$
- $r_1 =_{sc} r_2$ iff $s(r_1) = s(r_2)$ and $conf(r_1) = conf(r_2)$.

Using this partial order we can present the rules, e.g. as follows:

- Order the rules on support (use a die for a tie), from high to low.
- Per fixed support level, order the rules on confidence (again using dies), again from high to low.

Note, we can do this interactively, allowing the user to play with support and confidence levels, presenting only those rules that meet the currently set levels.
Clearly, we can mix the two approaches to rank results. For example:

- Order the rules by consequent.
- Per consequent order the rules by confidence and support.

While all these presentations create some order in the chaos of rules, they do not solve the problem of uninteresting results. Is there anything we can do?

What makes a rule interesting?

In principle a rule is interesting, if it gives useful (actionable) information. But is that something you can decide syntactically?

Lots of different measures [Bayardo, 1999] have been defined as an attempt to do just that, e.g. lift, interest, conviction, collective strength, gain, gini coefficients, entropy, χ^2, etc.... We will discuss a few of these.

Lift

The lift of an association rule tells us how much better the rule predicts the consequent than the random prediction:

$$lift(A \to C) = \frac{s(AC)/s(C)}{s(C)/|db|}$$

$$= conf(A \to C) \times \frac{|db|}{s(C)} \quad .$$

In other words, if a rule has a confidence of 0.9 while $s(C)/|db| = 0.2$, the lift of the rule is 4.5.

The term 'lift' comes from marketing. Suppose that you are interested in selling people C. If you random mail customers, you expect to sell to $s(C)/|db|$ customers. If you random mail A customers, you expect to sell to $conf(A \to C)$ customers.

So, if you mail the same number of people in both experiments, you expect to sell lift times as many in the second experiment. In other words, your sales figures are lifted by lift.

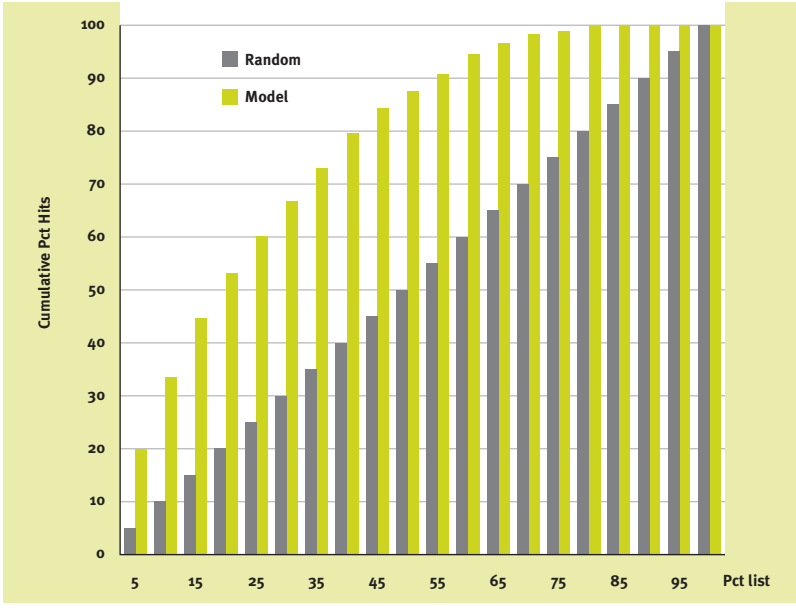


Figure 1
Lift: 5% of random hits have 5% of hits, but 5% of model-score ranked list have 21% of hits. Thus, Lift(5%)=21/5 = 4.2.

It is very well possible that $A \rightarrow C$ turns up, even if A and C are completely independent (the relation between probabilities and support and confidence is discussed in more depth later in this paper).

$$conf(A \rightarrow C) = P(C|A)$$
$$= \frac{P(AC)}{P(A)}$$
$$= \frac{P(A) \times P(C)}{P(A)} = P(C)$$

Interest and conviction

The interest of a rule measures how dependent A and C are by computing:

$$= \frac{P(A,C)}{P(A) \times P(C)} = \frac{s(AC)}{s(A) \times s(C)}$$

Interest is symmetric in A and C, it doesn't tell us whether A 'causes' C or the other way around. To get some feeling for this, it makes sense to look at A and C, in how far are these two independent (the 'attribute' C has value 1, when A has value 0 and vice versa):

$$conviction(A \rightarrow C) = \frac{P(A) \times P(\neg C)}{P(\{A, \neg C\})} = \frac{s(A) \times (|db| - s(C))}{s(A) - S(AC)}$$
$$= \frac{|db| - s(C)}{|db|(1 - conf(A \rightarrow C))}$$

The higher the conviction, the more often C occurs with A.

A high conviction doesn't mean a strong relationship. This strength is e.g. measured using the collective strength of an itemset, see [Agrawal, 2001]. A transaction is a violation of an itemset I, if the transaction contains some but not all items in I. The violation rate, denoted by $v(I)$, is the fraction of violations of itemset I over all transactions in the database. Given this notion, the collective strength of an itemset is defined as:

$$C(I) = \frac{1 - v(I)}{1 - E(v(I))} \times \frac{E(v(I))}{v(I)} \quad .$$

The expected number of violations $E(v(I))$ is computed assuming independence. Using such measures, we can weed out those rules that do not meet a certain threshold of interestingness. Combined with the schemes for presentation we discussed earlier, they should help the user to zoom in on the rules that are most interesting to him.

If we combine these two approaches, we can prove that the 'optimal' rules with respect to confidence and support are also 'optimal' with respect to measures such as lift and conviction. So, if we prune on these measures, we will retain those with 'optimal' support and confidence.

Generating less rules

Rather than post-processing the rules, one can also try to generate only the more interesting rules. There are two different approaches to this problem: within the confidence/support framework and outside this framework. That is, the latter class of algorithms does not select association rules on both confidence and support, but employs other interestingness measures directly from the start.
Within the confidence/support framework one of the approaches is to put extra constraints on the frequent sets. Two important concepts in this field are the maximal frequent itemsets and the closed frequent itemsets. The underlying idea of both concepts is that the set of all maximal/closed frequent itemsets represent all frequent itemsets, but is far smaller than the set of all frequent itemsets. Both are an example of what is also known as a condensed representation.
Maximal frequent itemsets are frequent itemsets such that none of their supersets is also frequent. Clearly, each frequent itemset is a subset of a maximal frequent itemset. MaxMiner [Bayardo, 1998] is an algorithm that directly mines maximal frequent itemsets from the database.
Closed frequent itemsets are itemsets that completely characterize their associated set of transactions. That is, a frequent itemset *A* is closed, if *A* contains all items that occur in all transactions in the support of *A*. The closed frequent itemsets form a basis for all frequent sets, see [Paquier, 1999] for more details as well as an algorithm for their discovery.
Another approach in the confidence/support framework is the multi-relational approach [Nijssen, 2001], (See Section 6.4.3). Although multi-relational mining simply implies that the database consists of several tables, one can easily use it to constrain the possible frequent itemsets. For example, by requiring that such itemsets should either cover a relation or not intersect that relation at all. The motivation for such an approach is that a relation describes some semantic entity, which can only exist completely or not at all.
The motivation for leaving the confidence/support framework is that support is not a very good interestingness measure. It implies that rules with a small support are not interesting. There are different approaches in this direction, all of these, however, rely to some extent on the independence of attributes just as most new interestingness measures.
An early example in this direction is the work of [Brin, 1997] mentioned above. Next to introducing χ^2 as an interestingness measure, the paper provides a direct algorithm for finding significant itemsets. However, support still plays a

role in the algorithm. The problem is that independence doesn't have an a priori-like pruning property. In other words, finding all significant itemsets is an exponential problem .

In [Castelo, 2001] we describe a way around this problem using Markov Blankets. More precisely, the Mambo *method* discovers all association rules $X \rightarrow Y$ such that Y is a singleton and X is a subset of a Markov Blanket of Y.

A Markov blanket of an attribute A is a minimal set of attributes MB such that A is conditionally independent of any other attribute B given MB.

The rationale for our approach is that A only depends on MB. More precisely, the rationale is that for any attribute $B \not\in MB \cup \{A\}$, we have that $P(A \mid MB,B) = P(A \mid MB)$. In other words if we know the state of MB, the state of B gives us no extra knowledge about the state of A.

So, rather than looking for frequent itemsets, our approach requires the search for Markov blankets. In Mambo, we use an MCMC algorithm for this problem. Where A Priori prunes on support, Mambo prunes on the posterior probability of the Markov blanket.

To keep our discussions simple, we assume that **X** is a singleton set denoted by X. For each attribute X, Mambo computes association rules using its (most probable) Markov blankets. Hence, it needs to know these blankets. The algorithm uses an oracle for this; hence the name: Maximum A posteriori Markov Blankets using an Oracle. There are various ways in which one can build an oracle for Mambo. We have chosen a Bayesian approach. The oracle provides a posterior distribution over the Markov blankets given the data. For any given Markov blanket, the higher the posterior, the more certain we are about the validity of that conditional independence. The reader should consult [Castelo, 2001] for full details on how this is implemented.

Understanding association rules

Consider:

		Loan = yes	**Loan = no**
Job = yes	house = yes	100	11
	house = no	100	89
Job = no	house = yes	50	39
	house = no	30	81

With a support level of 25%, the rule

Job = Yes → Loan = Yes

shows up, whereas

Job = *No* → *Loan* = *No*

does not, while the pair suggests the existing statistical correlation.
A related problem is when we have two rules $X_1 \rightarrow Y$ and $X_2 \rightarrow Y$. Do these rules describe different populations or not? For example, with a support level of 25% we get:

Job = *Yes* → *Loan* = *Yes*

House = *Yes* → *Loan* = *Yes*

While the underlying explanatory rule:

House = *Yes* ∧*Job* = *Yes* → *Loan* = *Yes*

does not show up!
The problem is that an association rule describes only one cell from a multi-dimensional contingency table. Other cells may contain information that is crucial towards understanding. And it is not given that these cells will show up as rules. One way to remedy this problem is by visualizing the contingency table using mosaic plots, as in Figure 2 [Hofmann, 2000].

A contingency table over *r* is basically a table of counts, in which each count denotes how often a given combination of attribute values occurs in *r*. The contingency table has a cell for each combination of attribute values of the participating attributes.
If we use $\sigma(v_{c_1}, \ldots, v_{c_k})$ as a shorthand for
$\sigma(A_{c_1} = v_{c_1} \wedge \ldots \wedge A_{c_k} = v_{c_k})$, the set of cells is given by:

$$\left\{\left(v, \left|\sigma(v)\right|\right) \middle| v \in D_{c_1} \times \ldots \times D_{c_k}\right\}$$

A useful aspect of this expression is that the *v*-values can be seen as the 'logical addresses' of the cells, whereas $|\sigma(v)|$ is the content of the cell with address *v*.
A mosaic plot visualizes each cell in a contingency table by a tile. The size of each tile is proportional to the count in the cell. That is, it is a set of tiles *t*:

$$\left\{(v,h,w) \middle| v \in D_{c_1} \times \ldots \times D_{c_k} \wedge h \times w = \left|\sigma(v)\right| / |r|\right\}$$

Clearly, this leaves the height (t.h) and the width (t.w) of tile t underspecified and the tiles can be all over the place. This is resolved by the construction of the plot.

Mosaic plots are constructed hierarchically, including the attributes one at a time in the user-specified order; say, by the list $[A_{m_1}, \ldots, A_{m_k}]$. We start with the initial Mosaic plot, which consists of one tile (the complete window) that represents the complete database.

Assume $[A_{m_1}, \ldots, A_i]$ have been included and we have $M_i = \{ (v, h, w) \mid v \in D_{m_1} \times \ldots \times D_i \}$. To include the attribute A_{i+1}, each tile $t \in M_i$ is split in subtiles, one for each value $val \in D_{i+1}$

The sizes of the subtiles should be proportional to the number of cases that fall in that specific intersection of attribute values. The easiest way to achieve this is by splitting the width or the height of the parent tile t.

This observation gives us two possibilities. Either we alternate between height-splitting and width-splitting, or we always split the width (or the height). The first option gives Mosaic plots, while the second option gives Double Decker plots.

One way to visualize an association rule $X \rightarrow Y$ is to combine all attributes involved in the left-hand-side selection X as *explanatory variables* and to draw them within one mosaic plot and to visualize the *response* Y by highlighting the corresponding categories in a bar chart.

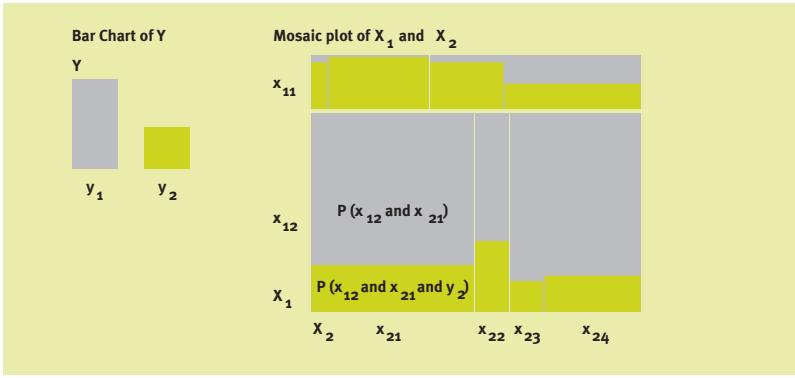


Figure 2
Mosaic plot.

Consider the association rule $X \rightarrow Y$ with $X = \{ A_{x_1} = v_{x_1} \wedge \ldots \wedge A_{x_k} = v_{x_k} \}$ and $Y = \{ A_y = v_y \}$. This association rule tells us something about the cells $(v_{x_1}, \ldots, v_{x_k})$, and $(v_{x_1}, \ldots, v_{x_k}, v_y)$ and their (relative) counts in the appropriate contingency tables. But the plot actually shows us more:

- it shows us all selections $\tilde{X}$ based on the attributes $A_{x_1}, \ldots, A_{x_k}$ for which the related rule $\tilde{X} \rightarrow Y$ holds;
- it shows us all selections $\tilde{X}$ based on the attributes $A_{x_1}, \ldots, A_{x_k}$ for which the rule $\tilde{X} \rightarrow Y$ holds nearly, as well as those selections for which $\tilde{X} \rightarrow Y$ doesn't hold at all;
- finally, by high-lighting different response values $A_Y = v_y$ we can see the interaction between the $A_{x_1}, \ldots, A_{x_k}$-values and the A_y value.

The same information can be seen from a double decker plot.

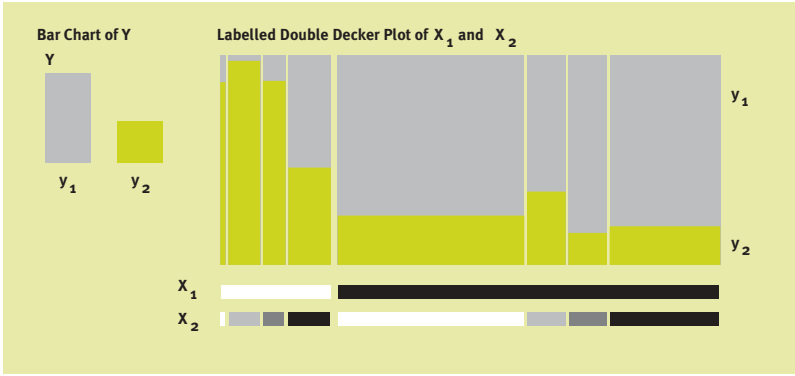


Figure 3
Double decker plot.

For a more concrete example, consider the next two plots.

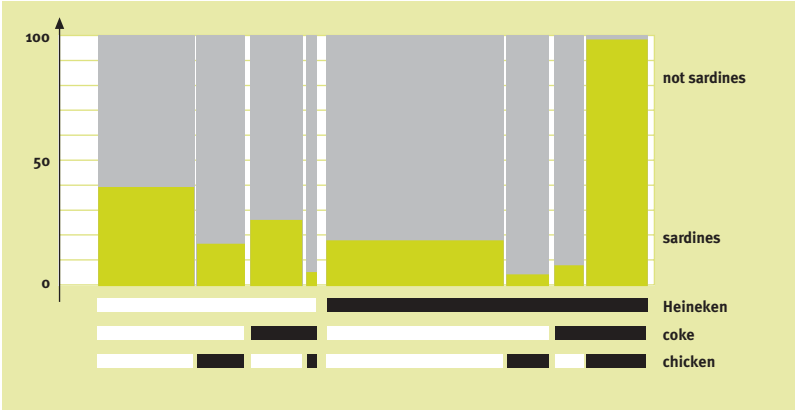


Figure 4
Double decker plot: The rule { Heineken, coke, chicken } → sardines is visibly a good rule. It has a relatively good support (it is a broad column) and has very good confidence (the sardines = yes part is very high), it is in fact far higher than for any other combination.

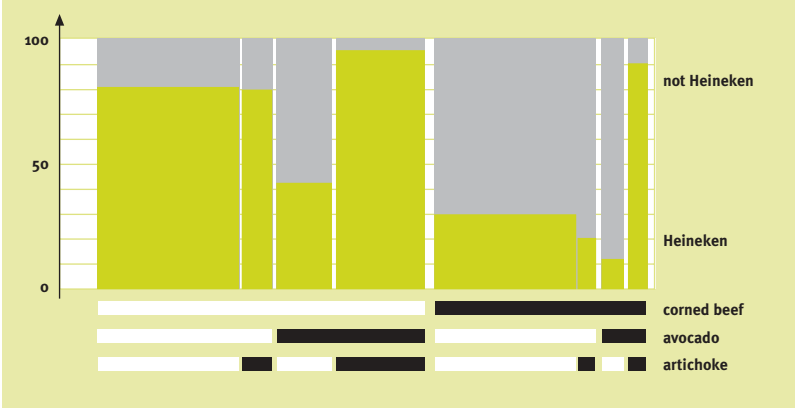
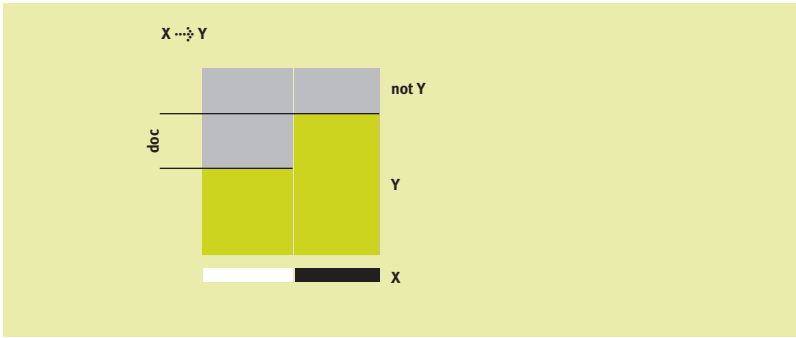


Figure 5
Double decker plot: The rule { corned-beef, avocado, artichoke } → Heineken is not a very good rule. Not only is the support small, but people that buy only artichoke and avocado have a higher probability of buying Heineken!

In using these plots, it is useful to consider an interestingness measure that one can see is:

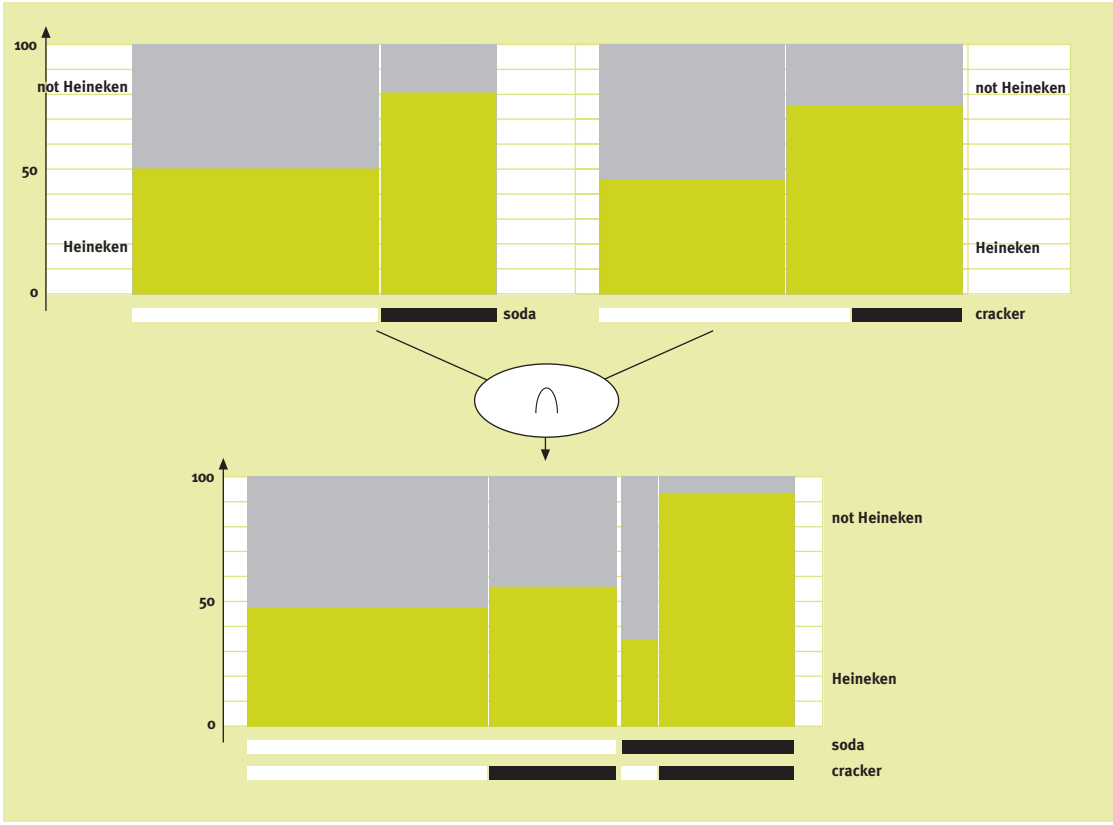
$$doc\,(X \to Y) := conf\,(X \to Y) - conf\,(\neg X \to Y).$$

Figure 6
Interestingness plot: doc (X → Y).



Double decker plots are also useful, when one tries to see whether two rules with the same right-hand side describe the same population or not. For this we use the intersection of plots. First consider:

Figure 7
Intersection of plots.



The real Heineken buyers are the people that buy both soda and crackers. In the next plot, however, we see that two other populations are distinct:

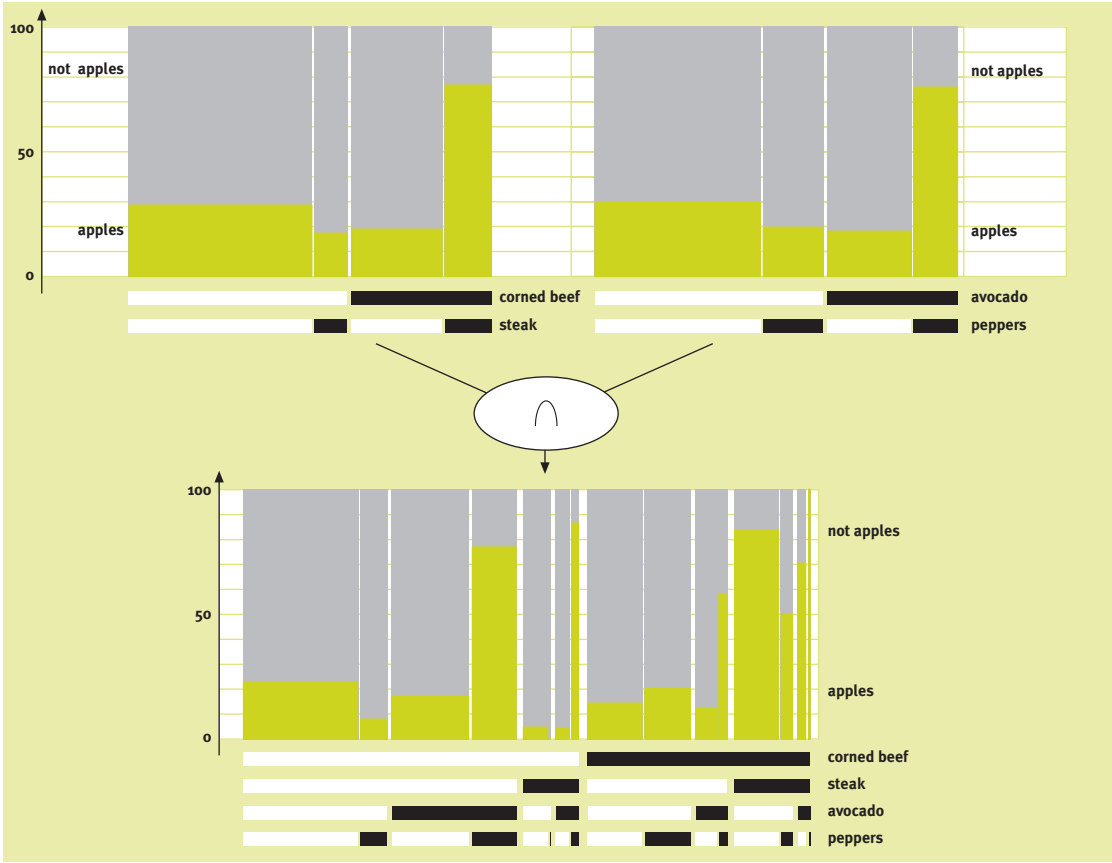


Figure 8
Identifying interesting populations with an intersection plot.

Association rules and categorical attributes

When the attributes are categorical rather than binary, we can, of course, still mine [Adamo, 2001; Han, 2001]. Rules are of the form:

$$A_1 = a_1 \wedge \ldots \wedge A_k = a_k \rightarrow B_1 = b_1 \wedge \ldots \wedge B_m = b_m$$

The support of an expression is defined by:

$$s(A_1 = a_1 \wedge \ldots \wedge A_k = a_k) = |\{ t \in db \mid \pi_{A_1 \ldots A_k} = (a_1, \ldots, a_k) \}|$$

The support and confidence of a rule are as defined before. Clearly, the A Priori property still holds, which means we can still use the A Priori Algorithm:

Compute the Frequent Sets
    Perform Level Wise Search
        Prune a Candidate Sets if one of
        its subsets is not frequent
Derive association rules

On the other hand, real valued attributes pose a problem. Either no expression like *salary* = 8234 will meet the minimal support or the number of resulting rules explodes and we will never get the complete answer.
The only way out of this predicament is to discretize real valued attributes. We can do this either in the data preparation phase or 'on the fly'.
If the real valued attributes are discretized before mining, we are back in the case of categorical attributes! The only difference is syntactical in that we write, for instance,

$age \in [18, 24]$

the rest remains the same.
Discretization on the fly means that we discretize a real valued attribute while mining for association rules. Usually this is done with some extra optimization criterion like lift:
Find (a) subset(s) of age such that:
- the lift is maximized;
- the minimal support is met (n times).

Note, that to retain the A Priori property, one can discretize only once!
There are (too) many possible intervals, therefore heuristics are often employed. In analogy to bump hunting (also known as subgroup mining), we can simply nibble small bits from either side of the interval such that:
- the optimization criterion rises maximally;
- the remaining interval has big enough support.

Stop when the optimization criterion no longer rises or the support becomes too small.
If a complete discretization is what you want, it may be better to use, e.g. a binary split algorithm:
- split the interval in two, such that the difference in the optimization criterion is maximal and the supports are big enough;
- recursively repeat the procedure on the sub-intervals until the criterion becomes constant or the support(s) become too small.

For real valued attributes we take care through discretization that values in the domain with little support are joined so that interesting rules actually can be found. For categorical values, no such thing happens and this might cause us to miss valuable information. For example, low minimal support is necessary to find rules like:

Computer = *PCG–Z600TEK* → *Camera* = *DSC–S70*

While a rule like:

Laptop Computer → Digital Still Camera

would be far more likely to appear.
Unfortunately, allowing arbitrary sets of values to be grouped causes some problems:

- There are $2^{|Dom|}$ possibilities.
- Most groupings make no sense at all.

These problems do not occur for real-valued attributes, because there is a natural grouping.
To overcome these problems, we will use hierarchies. A hierarchy is a set of sets. Each element set represents a level of the hierarchy:

- Level 1 has just 1 element.
- An element of level k 'belongs' to a unique element of level k-1.

There are a few questions we have to decide upon:

- Should all elements of a rule be on the same level of a hierarchy or not?
- Should we require the same support on each level or not?

The answer to such questions lies in the interpretation. Technically, we can only say that one should answer the questions coherently. Consider for example that the rule

Laptop Computer → Digital Still Camera

does not show up while the rule:

Computer = PCG–Z600TEK → Camera = DSC–S70

does, because we have a lower minimal support on a lower level. To me this seems inconsistent, but I can imagine uses for such unintuitive results. Whatever the answer to the questions, important is that we can simply use A Priori again.

Discovering hierarchies

We have seen that hierarchies of itemsets allow us to find strong associations that are too weak at lower levels. How do we get these hierarchies? There are two possible answers:

- They are a prime example of domain knowledge and, hence, are to be provided by the domain expert.
- Infer the hierarchies from the available data.

Hierarchies are similar to the dendographical trees of hierarchical clustering. In fact, if our distance function measures how (conceptually) different the items are, such a tree is a hierarchy in the sense we need it. Hence, we have to define the similarity of items and cluster. The similarity can be defined:

– just by the attributes themselves (internal measures);
– by their relations with other attributes (external measures).

When are two items *A* and *B* similar? If they are tend to be bought together:

$$d_{I_{sd}}(A,B) = \frac{\sigma\left((A=1 \wedge B=0) \vee (A=0 \wedge B=1)\right)}{\sigma(A=1 \vee B=1)}$$

$$= \frac{\sigma(A)+\sigma(B)-2\sigma(AB)}{\sigma(A)+\sigma(B)-\sigma(AB)}$$

In other words, the more often *A* and *B* are bought both when at least one of them is bought, the more similar *A* and *B*.

The strength of an association between items is measured by the confidence, so, we could also use that:

$$d_{I_{conf}}(A,B) = (1-conf(A \rightarrow B)) + (1-conf(B \rightarrow A))$$

$$= \frac{\sigma(B=0 \wedge A=1)}{\sigma(A=1)} + \frac{\sigma(A=0 \wedge B=1)}{\sigma(B=1)}$$

$$= 2 - \frac{(\sigma(A)+\sigma(B))\sigma(AB)}{\sigma(A)\sigma(B)}$$

It is easy to see that both $d_{I_{sd}}$ and $d_{I_{conf}}$ are metrics on the set of attributes. However, the similarity of two items, because they are often bought together is perhaps not what we are after. In fact, one would (almost) expect it the other way around. The fact that item *A* and *B* are very similar, means that *A* and *B* will not be bought together very often.

However, similarity doesn't just mean that the items are not often bought together. PC's and Remote Controls will not often be bought together, but PC's and RC's are hardly similar items. The similarity we are after is that customers bought item *A*, but could have bought item *B*. For example, Pepsi and Coke are very similar.

The key idea underlying the detection of this similarity is similarity with respect to all other products. Clients that buy Coke or Pepsi have similar buying behavior with respect to the other products. In other words, we should not look for a direct association between *A* and *B*, but for the similarity of their associations with the other products [Das, 1998].

Comparing the buying behavior of the *A* and *B* buyers with regard to all other products may be overdoing it a bit. Some items are simply not related to either *A* or *B* and it doesn't make sense to check such items. Moreover, the fact that *A* and *B* imply similar buying behavior for some products, but not for others gives a flexible notion of similarity.

We will only compare the behavior with respect to a probe set *P* of items. *P* captures the way in which we want *A* and *B* to be similar. If we want similar buying behavior for the items in *P*, we want the two relations:

$$r_A = \pi_p\left(Select\left(A=1\right)\left(r\right)\right)$$
$$r_B = \pi_p\left(Select\left(B=1\right)\left(r\right)\right)$$

to be similar. How do you determine the similarity of relations? Well, the relation r_A defines a probability distribution g_A on $\{0,1\}^P$ as follows:

$$x \in \left\{0,1\right\}^P : p\left(x\right) = \frac{\sigma_{r_A}\left(x\right)}{\left|r_A\right|} \quad .$$

Comparing two distributions can, e.g. be done with cross-entropy measures such as Kullbach-Leibler, however, this is computationally expensive for larger probe sets P. Hence, we have to look for cheaper (less well-based) solutions. In analogy with $d_{I_{conf}}$, we define:

$$d_{E_{conf},P}\left(A,B\right) = \sum_{D\in P}\left|conf\left(A\rightarrow D\right) - conf\left(B\rightarrow D\right)\right|$$
$$= \sum_{D\in P}\left|\sigma_{r_A}\left(D\right) - \sigma_{r_B}\left(D\right)\right| \quad .$$

in other words, we simply add the differences between the number of times *A* buyers bought *D* and *B* buyers bought *D*. If this total is low, *A* and *B* are deemed to be similar.

Clearly, this is not a metric, while *for A and B unequal*, $d_{E_{conf},P}(A, B)$ may very well be 0. In other words, $d_{E_{conf},P}$ is a pseudo metric. Fortunately, this doesn't matter for hierarchical clustering.

In principle, this measure is easy to compute. If we compute the association rules beforehand, we already know the values of $conf(A \rightarrow D)$ and $conf(B \rightarrow D)$, assuming *P* is well-chosen.

Similarly, if we 'join' *A* and *B* while clustering, we can compute

$$conf\left(A\vee B\rightarrow D\right) = \frac{\sigma\left(\left(A\vee B\right)\wedge D\right)}{\sigma\left(A\vee B\right)}$$

$$= \frac{\sigma\left(AD\right)+\sigma\left(BD\right)-\sigma\left(ABD\right)}{\sigma\left(A\right)+\sigma\left(B\right)-\sigma\left(AB\right)}$$

and we already know these σ values.

For two unrelated probe sets P_1 and P_2, one expects that the similarity measures d_{E_{conf},P_1} and d_{E_{conf},P_2} are completely unrelated and experiments show this to be true.
This means that one has to choose a 'good' probe set. What exactly defines a good probe set is a difficult question to answer, often requiring expert knowledge. Moreover, defining a probe set for soft-drinks should be easier than defining the exact hierarchy for soft-drinks.
What can be said, however, is that if $\{A_i\}_{i \in I}$ are the items one would like to cluster, then:

$$\forall D \in P \forall i \in I : A_i \rightarrow D$$

If only because this makes the computation of similarity cheap.
The discovered similarity of items has more uses than just for association rules. Potentially clients could buy either one. This is useful knowledge:

- If you ran out of one, you could suggest the other.
- If you discount one, you know that this (negatively) influences the sale of the other.

To be sure, however, you have to check that most or all of your clients are indifferent to the choice between the two items. How to find out? By classification, of course [Liu, 1998a, 1998b].

Associations and classification

If the table is divided into classes by a class attribute C, we can mine for association rules of the form:

$$X \rightarrow c_i$$

in which $c_i \in Dom(C)$ and X a condition on the other attributes. Clearly, such rules can be used to classify new tuples just as any other classifier system. Can we built a classifier system from association rules?
One of the important steps in classification tree building is pruning the tree. Trees are pruned to reduce the risk of overfitting. Similarly, some association rules may exhibit overfitting, if we want to use them for classification. Consider, e.g. the two rules:

$$X_1 \rightarrow c_1$$
$$\{X_1, X_2\} \rightarrow c_1$$

If the confidence of the second rule is hardly higher than that of the first rule, the second rule may exhibit overfitting.

There are many ways in which trees can be pruned:

- cost-complexity pruning;
- reduced error pruning;
- pessimistic pruning.

The first two require a test set, the latter estimates the error directly from the training data. This makes the integration of the last method into A Priori easier. If a leaf of a tree covers N tuples in the database and makes E errors in classifying those tuples, E/N can be seen as an estimate of the error inferred from N experiments with E successes.
The 'true' error rate can be higher, of course, we can bound the true error through confidence intervals:

$$P\left(m \notin CFI_{\delta}\left(N,E\right)\right) \leq \delta$$

If we denote the upper border of $CFI_{\delta}(N, E)$ by $U_{\delta}(N, E)$, we 'know' with 100–δ/2% certainty that the 'true' error rate will be less than $U_{\delta}(N, E)$.
In other words, we could prune if $U_{\delta}(N, E)$ is too large. How can we use this to prune association rules? Consider again the rules $X_1 \rightarrow c_1$ and $\{ X\text{–}1, X\text{–}2 \} \rightarrow C_1$. It seems reasonable to prune the latter, if it has a higher pessimistic error rate than the former. That is, we prune a rule r, if it has a sub rule r' with one attribute less that has a lower pessimistic error rate.

Association rules and time

One limitation of association rules as we have studied them is that they are only concerned with single moments of sale. While this makes sense for supermarkets, it makes far less sense for bookstores, video rentals/stores, etc.
Patterns that are useful for such stores are:

- If clients rent Rambo 1 this time, they are likely to rent Rambo 2 as one of their next movies.
- If clients have seen all Rambo movies, they tend to move on (up/down) to Jean-Claude van Damme movies.

To find such patterns, we should maintain databases that store sequences of client's sales [Mannila, 1995]. A sequence s is a list of itemsets:

$$s = [\, s_1, s_2, \ldots, s_n \,]$$

A sequence $[\, a_1, a_2, \ldots, a_n \,]$ is contained in the sequence $[\, b_1, b_2, \ldots, b_m \,]$ if

$$\exists i_1 < i_2 < \ldots < i_n : a_j \subseteq b_{ij}$$

For example:

$$[\{I_1\},\{I_2,I_3\}][\{I_1,I_2\},\{I_3\},\{I_1,I_2,I_3\}]$$

In a set of sequences, a sequence is maximal, if it is not contained in any other sequence.
The mining problem can now be formulated as follows. The database contains one sequence per customer (the customer sequences). The support of a sequence is defined by [Manilla, 1995]:

$$\sigma(s) = \frac{\left|\left\{c \in db \middle| s \prec c\right\}\right|}{|db|}$$

- Find all sequences S: $\sigma(s) \geq t$;
- Find all maximal frequent sequences.

And, again, we can do a level wise search:
- Checking whether one sequence is contained in another is costly and we have to do it often. Can we speed this up?
- We can compute all 'subsequences' of the customer sequences, then we only have to compare.
- We can make the comparison quicker by numbering all the sequences, since comparing numbers is far quicker.

To number the sequences, we need an injective function:

H: *Sequences* $\rightarrow N$

such a function is often called a (perfect) hash function. For example, if you want to hash words, you could use:

$$H(a_n \ldots a_1) = \sum_{j=1}^{n} n(a_j) 26^j$$

where $n(a_j)$ denotes the count of letter a_j. Note, here we assume that the count is always lower than 26.

Computing the maximal frequent subsequences is easy:
However, this is rather costly if we just want the maximal frequent sequences. Can we be smarter? The problem is that we generate many sequences that are not maximal. Computing the support of such sequences is a waste of time.
If we knew while constructing $F(i)$ the results of $F(i+k)$, we could skip counting all subsequences of elements of $F(i+k)$. The down-side of this is that computing

$F(i+k)$ before computing $F(i)$ is that we generate far too many candidates in $C(i+k)$, which is also costly. However, limiting k seems to do well in experiments.

Conclusions

Association rules are a powerful tool in the toolbox of the data miner. Association rules can be computed cheaply and provide useful insight in the data at hand. Although one tends to generate too many association rules, there are many useful techniques to filter the more interesting results from this flood. The concept of a frequent set and its monotonicity (the A Priori property) have been generalized and adapted to many other cases. One could easily fill a book on this topic; see [Adamo, 2001].

References

- Adamo, J-M. (2001). Data Mining for Association Rules and Sequential Patterns – Sequential and Parallel Algorithms. Springer Verlag
- Aggarwal, Yu. (2001). Mining Associations with the Collective Strength Approach. IEEE Transactions on Knowledge Discovery and Data Engineering
- Agrawal, R., T. Imilinski, A. Swami. (1993). Mining Association Rules between Sets of Items in Large Databases. Proceedings ACM-SIGMOD Conference
- Agrawal, R., R. Srikant. (1994). Fast Algorithms for Mining Association Rules. Proceedings 20th International Conference on Very Large Databases. VLDB94
- Bayardo. (1998). Efficiently Mining Long Patterns from Data. Proceedings ACM-SIGMOD Conference
- Bayardo, Agrawal. (1999). Mining the Most Interesting Rules. Proceedings KDD
- Brin, Motwani, Silverstein. (1997). Beyond Market Baskets: Generalizing Association Rules to Correlations. Proceedings ACM-SIGMOD Conference
- Castelo, Feelders, Siebes. (2001). MAMBO: Discovering Association Rules Based on Conditional Independencies. Proceedings 4th International Conference on Intelligent Data Analysis (IDA 2001)
- Das, Mannila, Ronkainen. (1998). Similarity of Attributes by External Probes. Proceedings KDD98
- Han, J., M. Kamber. (2001). Data Mining – Concepts and Techniques. Morgan Kaufman
- Hofmann, Siebes, Wilhelm. (2000). Visualizing Association Rules with Mosaic Plots. Proceedings KDD 2000
- Liu, Hsu, Ma. (1998). Integrating Classification and Association Rule Mining
- Liu, Hsu, Ma. (1998). Mining Association Rules with Multiple Minimum Supports. Proceedings KDD 1998
- Mannila, Toivonen, Verkamo. (1995). Discovering Frequent Episodes in Sequences. Proceedings KDD 1995

– Nijssen, Kok. (2001). Faster Association Rules for Multiple Relations. Proceedings 7th International Joint Conference on Artificial Intelligence (IJCAI'01)
– Pasquier, Bastide, Taouil, Lakhal. (1999). Discovering frequent closed itemsets for association rules. Proceedings 7th International Conference on Database Theory (ICDT)
– Proceedings KDD 1998
– Savasere, Omiecinsky, Navathe (1995). An Efficient Algorithm for Mining Association Rules in Large Databases. Proceedings 21st International Conference on Very Large Databases. VLDB95
– Toivonen, H. (1996). Sampling Large Databases for Association Rules. Proceedings 22st International Conference on Very Large Databases. VLDB96

6.2.13 Inductive logic programming

Maarten van Someren[1]

Introduction

Inductive Logic Programming (ILP) is not a particular method, but an approach to learning that includes a wide range of methods. The main characteristics of ILP are the use of logic[2] as a language for data and learning relational models. The ILP approach can be applied to any form of machine learning or data mining. In practice, methods have been developed for classification, modeling, prediction and adaptation.

Logic as language for data and models

Like other methods for machine learning and data mining, ILP methods take descriptions of observations and (optionally) prior knowledge about the goal of learning as input. In ILP, the input and also the resulting models are expressed in first order logic. ILP uses the syntax of formal logic, but what is more important, it allows the use of relations. First order logic is a language that is more expressive than standard attribute-value representations, because it includes relations. This makes it easy to express relations between objects, while generalizing over the nature of the objects themselves.

Suppose that we want to learn to recognize winning end game positions on the chessboard. One way of representing the chessboard is as a vector of which each dimension represents a field and the value of the piece on this field. Winning positions are characterized by positions of pieces relative to each other and not by absolute positions. For example, a threat to capture another piece is an important concept, but the precise form of the threat can vary enormously. A threat is less dangerous, if the threatened piece is covered by another piece. Which piece creates the threat, its precise position and the precise position of the threatened piece may vary widely. The relative positions of pieces are relevant. Expressed in terms of board positions of pieces, the relation 'X threatens to capture Y' is very complex and therefore difficult to learn. If relational concepts, like 'X on adjacent field to Y' or 'X on the same diagonal as Y' are used, then learning a relation like 'X threatens to capture Y' is much easier to learn. This is not just because concepts like 'X on adjacent field to Y' are good abstractions. A key property of the abstractions is that they can also abstract from, for example, which piece threatens another.

Another example, is a display of objects with spatial relations between them, like 'left-of', 'next-to', 'above'. Suppose that we want to learn under which conditions a product is likely to be bought. First-order logic makes it easy to express 'a cheap product next to an expensive product' as: next-to(X, Y) & cheap(X) & expensive(Y)'. This rule generalizes over the nature of the products.

1 Dr M.W. van Someren, maarten@swi.psy.uva.nl, Department of Social Science Informatics, The Universiteit van Amsterdam, The Netherlands
http://www.swi.psy.uva.nl/

2 See Section 6.2.1 for an overview of logic operators.

Attribute-based representations would need to enumerate the products and thus need much more complex models which are more difficult to find. Evidence for or against the relational rule comes from all sorts of products. Another advantage of the use of logic as a representation language is that learning can be extended to models that are organized as networks.

Powerful learning operations

There is a wide range of methods in ILP. Many are similar to methods from other approaches to data mining and therefore we summarize a few methods that are special to ILP. The main difference is that ILP methods learn models that include relations.

FOIL

FOIL is a method for constructing a set of rules for recognizing a target property or relation from a set of training examples. The method constructs rules of the form:

literal-1 & literal-2 & ... & literal-n → target-literal

for example:

next-to(X, Y) & cheap(X) & expensive(Y) → bought(X)

Training data for this rule are instances of next-to, cheap, expensive and bought, for example:

next-to(coffee-1, coffee-2)
cheap(coffee-1)
expensive(coffee-2)
unit size(coffee-1, 250)
unit size(coffee-2, 500)
organic(coffee-1)

These training data illustrate that the input data do not describe a single object with its properties, but several objects and their relations.
The rule above generalizes about the objects that appear in the instantiated facts and relations. The target can also be a relation. For example, we could learn the target relation 'prefers(X, Y)' instead of 'bought(X)'. FOIL constructs a set of rules for a target that expresses the conditions under which the target relation holds for a set of objects. The rule in our example means that if we have two objects X and Y and for these two objects next-to(X, Y), cheap(X) and expensive(Y) hold, then bought(X) (or 'prefers(X, Y)') will also hold.

FOIL constructs a set of general rules for a target concept that derives its target concept for all sets of objects for which the target concept is known to hold (in the training data). Each rule is learned by creating a general rule and then adding literals to exclude sets of objects for which the target relation does not hold. Candidate literals are terms (predicates with variables or constants as arguments), negative terms ('not expensive(Y)') or equalities that require variables to be equal. To avoid overly complex rules a criterion is needed for when there are not enough data to avoid rules that will not generalize well.

After constructing a single rule, FOIL removes the sets of objects covered by this rule from the training data and searches for sets of objects that satisfy the target concept, but are not yet covered by the rules that were constructed so far. A new rule construction process is started to make a general rule for the uncovered sets. The process ends when only a small number of uncovered sets of objects remains. To avoid overfitting a threshold is used to stop learning.

Inverted resolution and RLGG (Relative Least General Generalization)

A more drastic logical approach is to view learning as inverted deductive inference. We can formulate learning as 'finding a hypothesis H such that the following holds':

for all examples E: $H \cup E \Rightarrow$ target relation(E)

in other words: the hypothesis allows us for each example E to derive, if the target relation holds for E. This is especially interesting if we include (prior) background knowledge that can be used for the derivation. This gives: find H such that for all examples E: $B \cup H \cup E \Rightarrow$ target relation(E).

Abduction

ILP includes learning operators that are a forms of abductive reasoning: instead of deducing a conclusion from premises, premises are hypothesized from which certain facts could be deduced. Consider the following statements:

1 above(coffee1, coffee2)
2 cheaper(coffee1, coffee2)
3 bought(X) ← above(X, Y), cheaper(X, Y)
4 bought(coffee1)

Using logical deduction, we can use (1), (2) and (3) to deduce (4). Now suppose that we already know (4) and also (1) and (3). Abduction would now construct the missing premise (3) as a hypothesis. Another possibility is that from observation we know (1), (2) and (3). Abduction would now infer (3). Another possibil-

ity in this case would be the less general hypothesis (3a) *bought(coffee1)* ← *above(coffee1, coffee2), cheaper(coffee1, coffee2)*. It is easy to imagine that, if there are many known observations, abduction can produce a large number of hypotheses. Once a hypothesis is abductively inferred, it can itself be used in further abductive inferences.

To control the learning process, ILP methods rely on general principles like simplicity (a general theory allows specific observations to be dropped; learning can thus be guided by the goal of finding simple theories that are still consistent with the observations) and prior knowledge. E.g. systems usually allow the user to specify constraints on the form of hypotheses.

Some points:

- This is a very computationally expensive and explosive method.
- It is guaranteed to find the most specific clause under theta-subsumption[3].
- Hypotheses can be recursive; this may produce non-terminating programs.

ILP methods may create large search spaces

The main strengths of ILP are that it gives a framework for analyzing and designing representation forms for observations, models and prior knowledge. ILP gives a natural way to address issues like learning relational models, including background knowledge and learning recursive models.

A very general notion of 'more general than':

'T1 is more general than T2' ⇔ 'T1 ¬ T2'
so: learning is a kind of inverse deduction,
compare:

1 T1 ¬ T2
2 from T2 learn T1 such that (1) holds

The main weakness of the approach is that the use of expressive representations implicates very large numbers of possible models, while logic does not provide much guidance in organizing the construction of hypotheses in the face of this complexity. For example, the use of full first order logic[4] would make the task of prediction subject to the halting problem. First-order logic is not able to decide and could lead to infinite computation. This makes it necessary to combine logical representations and methods with bias that restricts the search.

3 An operator reducing computational cost by allowing incompleteness.

4 See Sections 6.2.1 and 2.2.4.

Extensions beyond Horn clause logic

The use of first-order logic as a representation language brings with it both the strengths and weaknesses of logic. The main strength is the expressiveness of

the representation language, which may include recursive expressions. The main weaknesses are complexity and unmanageability of the learning process and limited methods for numerical, stochastic and dynamic domains. However, recent results show that ILP, like machine learning in general, can be extended to include methods from numerical modeling, a probabilistic version of logic and models of dynamic changes in the world that is modeled.

Some basic concepts and terminology

(Relative) Least General Generalization ((R)LGG): given a set of clauses for a target predicate, find the least general set of clauses that can deduce all positive instantiations of the target predicate and no instantiations known to be false (or up to a thresholds based on a statistical criterion).
Inverted resolution: learning operator that inverts the resolution operator. It generates one or more theorems that would allow a given theorem to be deduced by resolution.
Abduction: form of reasoning that infers knowledge that would allow given observations to be derived or explained. Inverted resolution is a form of abduction.
Relational learning: learning about structured descriptions that involve multiple objects, properties of these objects and relations between them.

Data mining terminology	**ILP terminology**
Training examples	Logical axioms (describing data)
Model	Theorems
Model reproduces data	Model theorems LOGICALLY ENTAIL data axioms
Constructing model	Inductively inferring model
Prior knowledge	Logical axioms (for prior knowledge)
Language bias	Declarative bias

Example applications

Learning to predict the chemical structure of diterpenes from magnetic resonance spectra

Diterpenes are organic compounds of low molecular weight with a skeleton of 20 carbon atoms. They are of significant chemical and commercial interest, because of their use as lead compounds in the search for new pharmaceutical effectors. The interpretation of diterpene 13C NMR spectra normally requires specialists with detailed spectroscopic knowledge and substantial experience in natural products chemistry, more specifically knowledge on peak patterns and chemical structures. Given a database of peak patterns for diterpenes with known structure, we apply several ILP approaches to discover correlations

between peak patterns and chemical structure. Performance close to that of domain experts is achieved, which suffices for practical use. For more information see [Džeroski, 1998a].

Discovering properties of chemicals

A number of studies performed by Computing Lab, Oxford University in collaboration with others address the problem of discovering unknown properties of chemicals. One such property is toxicity. For research and development of new medicine and also for other chemicals it is important to know in advance, if the chemical will be toxic. One of the purposes of these studies is to construct a model that can predict from their chemical composition and structure, if new chemicals will be toxic. This can be used to avoid expensive testing or even chemical engineering to construct the substance.

As input were used descriptions of the chemical structures of chemicals and their known toxicity. ILP methods were used, because often the structure of organic chemicals is important and this needs generalization over structures besides values of attributes of chemicals.

For more information see [Toxicology, 2001; Srinivasan, 1999; King, 1995].

Learning to recognize critical road sections

When a road is jammed, the cause can be elsewhere. A research group at Univ. Leuven used machine learning (ILP) to learn from accident descriptions and knowledge about the road system, to recognize the probable cause of a traffic problem [Džeroski, 1998b].

References[5]

- Džeroski, S., S. Schulze-Kremer, K. Heidtke, K. Siems, D. Wettschereck, H. Blockeel. (1998a). Diterpene Structure Elucidation from 13C NMR Spectra with Inductive Logic Programming. Applied Artificial Intelligence **12**:363-384
- Džeroski, S., N. Jacobs, M. Molina, C. Moure. (1998b). ILP Experiments in Detecting Traffic Problems. In: C. Nédellec, C. Rouveirol. (ed.). Proceedings of the 10th European Conference on Machine Learning (ECML'98). LNAI **1398**:61-66. Springer Verlag, http://www.cs.kuleuven.ac.be/cwis/research/dtai/publications/1998-N.shtml
- King, R.D., A. Srinivasan, M.J.E. Sternberg. (1995). Relating Chemical Activity to Structure: an Examination of ILP Successes. New Generation Computing **13**:411-433
- Srinivasan, A., R.D. King. (1999). Feature Construction with Inductive Logic Programming: a Study of Quantitative Predictions of Biological Activity Aided by Structural Attributes. Data Mining and Knowledge Discovery **3**:37-57
- Toxicology Challenge. (2001). About a Competition on Learning Predictive Models. http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/

5 For further reading, the CD-rom contains an updated summary of the book Inductive Logic Programming: Techniques and Applications, by N. Lavrač and S. Džeroski. Ellis Horwood, New York, 1994. The full book is also included, with kind permission of the authors.

6.2.14 Rule induction by bump hunting

Ad Feelders[1]

Introduction

In this section we discuss a class of algorithms that tries to find regions in the input (attribute/feature) space with relatively high (low) values for the target variable. The regions are described by simple rules of the type

if: condition-1 and ... and condition-n
then: estimated target value.

There are many problems where finding such regions is of considerable practical interest. Often these are problems where a decision maker can in a sense choose or select the values of the input variables so as to optimize the value of the target variable.
Consider, for example, the problem of loan acceptance faced by a bank. Obviously, the bank would prefer to grant loans to people with a low risk of defaulting, and reject applicants with a high risk. It is assumed (and evidence shows) that the risk of defaulting depends on characteristics of the applicant, such as income, age, occupation, and so on. Now the bank may have collected data in the past concerning the characteristics of accepted applicants together with the outcome of the loan (defaulted or not). Such data may be used to find groups of applicants with a low probability of defaulting, which clearly is valuable information in deciding whether or not to accept future applicants.
Other applications are for example, the identification of interesting market segments and industrial process control.

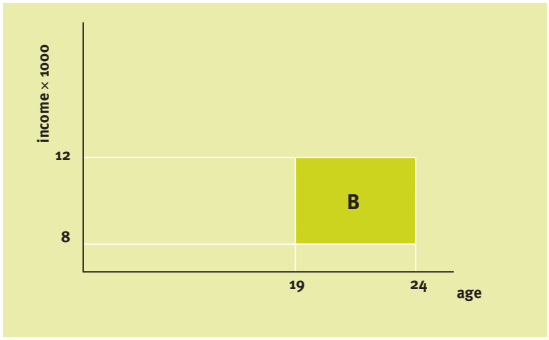
Shape of the regions

The regions we are looking for must have 'rectangular' shape, hence we call them boxes. Figure 1 shows an example of a box defined on two numeric variables, where persons with an age between 19 and 24, and income between 8,000 and 12,000 fall into the box. Figure 2 shows an example of a categorical box, where females who are single or married fall into the box.
Boxes may also be defined on combinations of numeric and categorical variables. An important property of a box is its support, which is the fraction of points from the data set that fall into the box.

Covering

In bump hunting it is customary to follow a so-called covering strategy. This means that the same box construction (rule induction) algorithm is applied sequentially to subsets of the data. The first box is constructed on the entire

1 Dr A.J. Feelders, ad@cs.uu.nl,Utrecht University, The Netherlands Institute of Information & Computing Sciences, Utrecht, The Netherlands, http://www.cs.uu.nl/



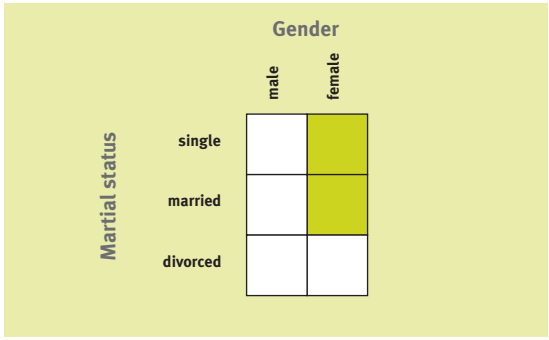


Figure 1 (left)
Numeric box.

Figure 2 right)
Categorical box.

data set. For the construction of the second box, we remove the data points that fall into the first box. For the construction of the third box, we remove the data points that fall into the first or second box, and so on.

In computing the support of a box, we count the data points that fall into that box (but not into any of the previous boxes) and divide this count by the number of observations of the entire data set (not just the data we used to construct the current box). Box construction continues until there is no box in the remaining data with sufficient support and sufficiently high target mean.

Box construction (rule induction)

Given the data (or a subset of the data), the goal is to find a box within which the target mean is as large as possible. It is not feasible to simply consider all possible boxes and pick the one with the highest target mean, so usually some kind of heuristic search is performed to find a good box. There are different strategies to come up with good boxes, and these strategies are discussed in more detail on the CD-rom. Here we give an outline of the basic ideas and some examples.

As already mentioned we start out with all the (remaining) data. Next we select a subset of the data on the basis of a condition on one variable. From all possible selections we choose the one that leads to the highest target mean within the selected group (provided that the group has enough support). After the first step we continue with the selected group and again look for the best selection we can make on a single variable. This process continues until the support of the remaining group drops below the minimum threshold set by the user.

The search strategy described above is called hill climbing, because at each step we continue with the single best selection we can make at that point. Alternatively, we could look at more than one selection, say the best 4. This is called a beam search, and the number of selections we consider at each level in the search is called the beam width. We may find better rules with a beam search, because, for example, the selection that only ranked third at the first step may later turn out to lead to a better group than the selection that ranked first.

We give an illustration using a data set concerning 1,519 households drawn from the 1980-1982 British Family expenditure survey. For each household we have data on the budget share spent on different expense categories (e.g. food, clothing, alcohol, and so on), as well as data on total household expenditure (totexp), total net household income (inc), age of household head (age), and number of children in the household (nk).
With bump hunting we may for example look for groups of households that spend a relatively large share of their budget on food. On average the households in the sample spend about 36% of their budget on food.
We have analyzed this data set with PRIM [Friedman, 1999], a bump hunting algorithm developed by Friedman and Fisher. PRIM uses a hill climbing strategy to find good groups. The best rule found on the entire data set is:

if totexp < 45 and age > 33 and inc < 135
then wfood = 58%

where wfood denotes the budget share spent on food. This group has a support of about 1% (the minimum support threshold). Next the data points that match this rule are removed from that training data, and we proceed with the remaining data to find the second rule.
As a second example, we analyze a dataset concerning car insurance policies, where the target variable indicates whether the insurant claimed or not. We are interested in high risk groups, so we look for subgroups in the data that have an above average tendency to claim. Figure 3 shows the result of a beam search by Data Surveyor [Holsheimer, 1996; Siebes, 1995], with the beam width set to three, and minimum support set to 1%. The leftmost node in the Figure represents the entire data set, which has exactly 50% claimers and 50% non-claimers (the data was selected to achieve this balance). The confidence interval for the probability of claiming is indicated between square brackets. Directly to the right of the leftmost node, the three best boxes (subgroups) are displayed. Data Surveyor only considers boxes that have a significantly higher target mean than their parent box. When the confidence intervals are non-overlapping, the difference is considered to be significant. The number of data points in a box is given next to the confidence interval. Note that at level 3 in the search we only have two boxes, even though the beam width was set to three. This is simply, because there was no other box with a non-overlapping confidence interval and support higher than 1%.
From Figure 3 we can read that the most risky group found are young people who drive expensive lease cars. It is very likely that the men of this group are more risky than the women, but their support is too small.

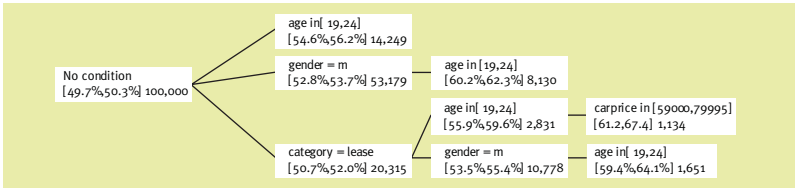


Figure 3
Example of beam search: beam width=3 and minimum support=1%.

Using the results of bump hunting

We have discussed the covering strategy of bump hunting: find the best rule (box) on the complete dataset, remove the data points that fall into that box, and proceed to find the best rule on the remaining data. The end result is a list of rules where each rule specifies a box on the input variables and an estimated target mean in that box.

The most obvious use of such a list of rules is for the prediction of the target value of a new case with unknown target, or the selection of cases with high predicted target value. For example in credit scoring we may want to use the list of rules to select applicants with a low probability of defaulting. When a new applicant arrives, we collect the relevant data on the input variables, and proceed as follows. We look whether the applicant matches the first rule in the list. If so, we look at the associated probability of defaulting, which is presumably very low for the first rule. The applicant would be accepted in that case. If he or she does not match the first rule, we proceed to the second rule, and so on. If the applicant does not match any of the rules, or we consider the probability of defaulting associated with the first matching rule to high, the applicant is rejected.

Summary

Bump hunting algorithms are very suited, when we are interested in finding groups in the data that have a particularly high (or low) value for the target variable. The groups are typically described by the conjunction of a number of simple conditions, each condition based on a single input variable. This has the advantage that the individual rules are easy to interpret.

References

- Friedman, J.H., N.I. Fisher. (1999). Bump-Hunting in High-Dimensional Data. Statistics and Computing **9**:123-143
- Holsheimer, M., M. Kersten, A. Siebes. (1996). Data Surveyor: Searching the Nuggets in Parallel. In: U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, U. Uthurusamy. (eds.). Advances in Knowledge Discovery and Data Mining. pp447-467. AAAI Press
- Siebes, A. (1995). Data Surveying: Foundations of an Inductive Query Language. In: U. Fayyad, U. Uthurusamy. (eds.). Proceedings of the First International Conference on Knowledge Discovery and Data Mining. pp269-274. AAAI Press

Introduction to evolutionary computing

A.E. Eiben[1]

Abstract

Evolutionary computing is an exiting development in computing science. It amounts to building, applying and studying algorithms based on the Darwinian principles of natural selection. In this paper we briefly introduce the main concepts behind evolutionary computing. We present the main components shared by all evolutionary algorithms (EA) and sketch differences between different types of EAs. Then we consider the main achievements (including the major application areas ranging from optimization, modeling and simulation to entertainment) and some important challenges of evolutionary computing.

Motivation and history

Evolutionary computing has its origins in biology and automated problem solving. Developing automated problem solvers (that is, algorithms) is one of the central themes of mathematics and computer science. Similarly to engineering, where looking at Nature's solutions has always been a source of inspiration, copying 'natural problem solvers' is a stream within this discipline. In particular, one can ask the following question: What is the most powerful problem solver in the universe?

Two answers are straightforward:

– The human brain that created 'the wheel, New York, wars and so on'[2].
– The evolutionary process that created the human brain.

Trying to design problem solvers based on these answers leads to the fields of neurocomputing, respectively evolutionary computing. The fundamental metaphor of evolutionary computing relates natural evolution to problem solving in a trial-and-error (also known as generate-and-test) fashion.

Table 1
The basic metaphor of evolutionary computing.

Evolution		Problem solving
Environment	⟷	Problem
Individual	⟷	Candidate solution
Fitness	⟷	Quality

In natural evolution, a given environment is filled with a population of individuals that strive for survival and reproduction. Their fitness — determined by the

1 Prof Dr A.E. Eiben, gusz@cs.vu.nl, Computational Intelligence Group, Faculty of Sciences, The Vrije Universiteit Amsterdam, Amsterdam, The Netherlands, http://www.cs.vu.nl/ci/

2 After Douglas Adams' Hitch-Hikers Guide to the Galaxy.

environment — tells how well they succeed in achieving these goals, i.e. it represents their chances to live and multiply. In the context of a stochastic generate-and-test style problem solving process we have a collection of candidate solutions. Their quality — determined by the given problem — determines the chance that they will be kept and used as seeds for constructing further candidate solutions.

Surprisingly enough, this idea of applying Darwinian principles to automated problem solving dates back to the forties, long before the breakthrough of computers, [Fogel, 1998]. As early as in 1948 Turing proposed 'genetical or evolutionary search' and already in 1962 Bremermann actually executed computer experiments on 'optimization through evolution and recombination'. During the sixties three different implementations of the basic idea were developed at three different places. In the USA Fogel introduced evolutionary programming [Fogel, 1966; Fogel, 1995], while Holland called his method a genetic algorithm [Holland, 1992; Goldberg, 1989]. In Germany Rechenberg and Schwefel invented evolution strategies [Schwefel, 1995]. For about 15 years these areas developed separately; it is since the early nineties that they are envisioned as different representatives ('dialects') of one technology, that was termed evolutionary computing [Bäck, 1996; Bäck, 1997; Eiben, 1998; Michalewicz, 1996]. It was also in the early nineties that a fourth stream following the general ideas has emerged — Koza's genetic programming [Koza 1992; Banzhaf, 1998]. The contemporary terminology denotes the whole field by evolutionary computing, or evolutionary algorithms, and considers evolutionary programming, evolution strategies, genetic algorithms, and genetic programming as subareas.

What is an evolutionary algorithm?

The common underlying idea behind all these techniques is the same: given a population of individuals, the environmental pressure causes natural selection (survival of the fittest) and hereby the fitness of the population grows. It is easy to see such a process as optimization. Given an objective function to be maximized, we can randomly create a set of candidate solutions and apply the objective function as an abstract fitness measure (the higher the better). Based on this fitness, some of the better candidates are chosen to seed the next generation by applying recombination and or mutation. Recombination is applied to two selected candidates (the so-called parents) and results in one or two new candidates (the children). Mutation is applied to one candidate and results in one new candidate. Applying recombination and mutation leads to a set of new candidates (the offspring) that competes — based on their fitness — with the old ones for a place in the next generation. This process can be iterated until a solution is found or a previously set time limit is reached. Let us note that many components of such an evolutionary process are stochastic. So is selection,

where fitter individuals have a higher chance of being selected than those who are less fit, but typically even the weak individuals have a chance of becoming a parent or of survival. For recombination of information of two individuals it holds that the choice on which pieces of information will be exchanged is random. Similarly for mutation, the pieces that will be mutated within a candidate solution and the new pieces replacing the old ones are chosen randomly. The general scheme of an evolutionary algorithm can be given in the Inset.

Inset: The general scheme of an evolutionary algorithm
Initialize population with random candidate solutions
Compute fitness of each candidate
WHILE not stop DO
Select parents
Recombine pairs of parents
Mutate the resulting offspring
Compute fitness of new candidates
Select survivors for the next generation
OD
Return best individual of the generation as **solution**.

Let us note that this scheme falls in the category of generate-and-test, also known as trial-and-error algorithms. The fitness function represents a heuristic estimation of solution quality and the search process is driven by the variation operators (recombination and mutation creating new candidate solutions) and the selection operators. In fact, there are two forces behind an evolutionary process:

1 One force is increasing diversity. This is realized by the variation operators recombination and mutation that create new points in the search space. This force represents a push towards novelty.
2 The other force is decreasing diversity. This is realized by selection operators for parent selection and survivor selection. This force represents a push towards quality.

Evolutionary algorithms (EA) are distinguished within in the family of stochastic generate-and-test methods by the following features:

– EAs are population based, i.e. they process a whole set of candidate solutions.
– EAs mostly use recombination to mix information of two candidate solutions into a new one.

The aforementioned 'dialects' of evolutionary computing follow the above general outlines and differ only in technical details. For instance, the representation

of a candidate solution is often used to characterize different streams. Traditionally, the candidates are represented by (i.e. the artificial DNA encoding a solution has the form of) bit-strings in genetic algorithms , real-valued vectors in evolution strategies, finite state machines in evolutionary programming and trees in genetic programming. The borders between the different streams are, however, diminishing. Clearly, the recombination and mutation operators working on such candidates must match the given form. That is, for instance in genetic programming the recombination operator works on trees, while in GAs it operates on bit-strings. It is important to note that selection takes only the fitness information into account, hence it works independently from the actual representation. Differences in the commonly applied selection mechanisms in each stream are therefore rather a tradition than a technical necessity.

Achievements of evolutionary computing

One of the most important achievements of evolutionary computing (and related fields) is that it supplies evidence for the practicability of 'intelligence without reasoning', or in other words 'problem solving without intelligence'. These terms and the phenomena they point to go further than traditional artificial intelligence, which is based on modeling human reasoning processes. Computational intelligence[3] shows that there is a working alternative to this AI approach.

Successful applications form an important part of the evidence for the power of evolutionary computing. Although it is often stressed that an evolutionary algorithm is not an optimizer in the strict sense, optimization problems form the most important application area. Especially, on hard combinatorial optimization problems (NP-hard, NP-complete)[4] evolutionary algorithms often outperform other, traditional methods. This holds not only for academic benchmark problems, e.g. graph coloring, or 'satisfiability', but also for problems in practice. For instance, the Napier University in Edinburgh, Scotland, uses a genetic algorithm for making the complete teaching and exam time table for all faculties of the university. Machine learning and modeling tasks form another important field where EAs are successfully applied. A Dutch example is the data mining system Omega from Cap Gemini based on genetic programming. The system, applied in banking, insurance, telecom, etc. has been a commercial success for years. Recent developments show a growing success of evolutionary design applications. Industrial design, such as designing jet engine parts, satellite constructions or automobile components by evolutionary means proves to be more and more successful. The technique is used by, for instance, NASA, Daimler-Chrysler and British Telecom. Immediately related to this area we find evolutionary art, where the designed objects are pieces of art in a given art form, e.g. pictures or music. Typically, the fitness is based on so-called subjective selection,

3 Usually 'defined' as the collective name for evolutionary computing, fuzzy computing, and neurocomputing.

4 Measures of computational complexity. NP: Guessed solutions for a problem with size n can be checked in time t= $O(n^k)$ (k= constant). NP-hard, problem as hard as or harder than any problem in NP. NP-complete, problems which are both NP and NP-hard. NP stands for Nondeterministic Polynomial time.

based on the users appreciation of the pieces in a given population. Variation operators are then applied only to the user-selected pieces to ensure that the next generation will be 'nicer' than the present one. The art museum 'Het Gemeentemuseum' in The Hague in The Netherlands featured an Escher-evolver *in vivo* as part of the large Escher overview exhibition in 2000.

Besides such applications directed to problem solving, evolutionary processes are also used for simulation purposes. The areas of artificial life, evolutionary economy, computational societies typically simulate certain phenomena by an evolving system. Such studies are often directed to what-if like questions, rather than to seeking solutions for a given problem.

Challenges of evolutionary computing

The theory — or rather, the lack of theory — of evolutionary computing remains one of the greatest challenges. From the early days on there have been mathematical analyses of certain aspects of evolutionary algorithms such as the well-known schema theorem of genetic algorithms, convergence rates of evolution strategies, or Markov chain analysis of EA's in general. The results, however, have been questioned. Not for their technical correctness, but their relevance, i.e. their explanatory and predictive power concerning the behavior of an evolutionary algorithm. Present theory forming often borrows machinery from other areas, like quantitative genetics or statistical physics, [Eiben and Rudolph, 1999], while some view a genetic algorithm as a 'purely mathematical object' and try to analyze it in a 'purely mathematical way' [Vose, 1999].

A second great challenge which is being taken up by more and more researchers in evolutionary computing is related to the evolution of the evolution, or optimizing the optimizer. Technically, it concerns the design of a high performance algorithm setup for a given problem or problem class. Since EA's have quite a few parameters, like for instance the population size, the frequency of mutations, or the strength of selection, tuning them to a problem can be time-consuming[5]. However, many results indicate that the evolution mechanism is robust enough to calibrate itself, that is, adjusting its own setup to a given problem while solving the problem, cf. [Eiben et al, 1999]. Developments into this direction will expectedly lead to a new type of self-adjusting algorithms having a great impact on practice and requiring a new type of theory.

5 Earlier claims that EAs are not very sensitive for their parameter settings have been revised. While they can really deliver satisfactory results without much tuning, appropriate parameter choices are important for top performance.

Closing remarks

Natural evolution can be considered as a powerful problem solver achieving Homo Sapiens from the primordial soup in only a couple of billion years. Computer-based evolutionary processes can also be used as efficient problem solvers for optimization, constraint handling, machine learning and modeling

tasks. Furthermore, many real-world phenomena from the study of life, economy, and society can be investigated by simulations based on evolving systems. Last but not least, evolutionary art and design form an emerging field of applications of the Darwinian ideas. We expect that computer applications based on evolutionary principles will gain popularity in the coming years in science, business, and entertainment. Computer evolution also remains the subject of further research. The coming years are facing challenges, among others in theory and in the novel area of self-adjusting algorithms.

References

- Bäck, T. (1996). Evolutionary Algorithms in Theory and Practice. Oxford University Press
- Bäck, T., D. Fogel, Z. Michalewicz. (eds.). (2000). Handbook of Evolutionary Computation. Institute of Physics Publishing, Bristol and Oxford University Press. 2nd edition
- Banzhaf, W., P. Nordin, R.E. Keller, F.D. Francone. (1998). Genetic Programming: An Introduction. Morgan Kaufmann
- Eiben, A.E., Z. Michalewicz (eds.). (1998). Evolutionary Computation. IOS Press
- Eiben, A.E., G. Rudolph. (1999). Theory of Evolutionary Algorithms: a Bird's Eye View. Theoretical Computer Science **229**:3-9
- Eiben, A.E., R. Hinterding, Z. Michalewicz. (1999). Parameter Control in Evolutionary Algorithms. IEEE Transactions on Evolutionary Computation **3** (2):124-141
- Fogel, L.J., A.J. Owens, M.J. Walsh. (1966). Artificial Intelligence through Simulated Evolution. J. Wiley, New York
- Fogel, D.B. (1995). Evolutionary Computation. IEEE Press
- Fogel, D.B. (1998). Evolutionary Computation: the Fossil Record. IEEE Press
- Goldberg, D.E. (1989). Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley
- Holland, J.H. (1992). Adaptation in Natural and Artificial Systems. MIT Press
- Koza, J.R. (1992). Genetic Programming. MIT Press
- Michalewicz, Z. (1996). Genetic Algorithms + Data Structures = Evolution Programs. Springer Verlag. 3rd edition
- Schwefel, H.-P. (1995). Evolution and Optimum Seeking. J. Wiley, New York
- Vose, M.D. (1999). The Simple Genetic Algorithm: Foundations and Theory. MIT Press

Evolutionary algorithms for data mining

Robert E.J. Keller[5]

Evolutionary algorithms and genetic programming

For real-world problems — i.e. practical decision problems — e.g. scheduling, routing, and network layout, to name but a few of hundreds of problem classes, one is interested in 'good' solutions. Especially, an optimal solution is highly desirable, since even tiny differences in solution qualities may represent vast financial differences on a large industrial scale. The problems come in different sizes, that is, their numbers of decision variables differ. For many real-world problems, all those known deterministic algorithms which certainly deliver an optimal solution, e.g. enumeration, feature a run-time consumption that grows exponentially over the problem size. Usually, the problem sizes of these problems do not remotely allow for the use of exact algorithms. To make things worse, the mathematical representation of a typical real-world problem is often non-differentiable, non-convex and multimodal, which renders the use of many classic optimization algorithms impossible.

Thus, the use of stochastic algorithms guided by heuristics is the only option, when being faced with a problem instance whose size renders the use of deterministic algorithms infeasible. Evolutionary algorithms are typical examples of such algorithms as they employ algorithmic metaphors of random processes and, as heuristics, principles which are effective in organic evolution.

When starting, evolutionary algorithms create a population of individuals which represent potential solutions, that is, points in the search space of the problem. Preferably from the better of these individuals, changed or copied individual offspring are generated which themselves may serve as parents, depending on their quality which is also called fitness. These parents lead to new offspring and so forth, which makes an evolutionary algorithm an open-ended search process that, in practice, may be ended, when it meets a termination criterion. Those criteria may be, for instance, the full consumption of available computing resources or the identification of a solution that satisfies the decision maker.

There are 'generational' evolutionary algorithms that produce a distinct set of offspring individuals from a distinct set of parental individuals, and such a set is called generation. The offspring generation then becomes the next parental generation, which starts the next cycle.

The variation of the individuals is mainly performed by two types of genetic operators called recombination and mutation whose operation partially depends on random processes. The decision as to which individual may become input to a genetic operator is made by a selection operator that mostly chooses individuals with an above-average quality. The quality of an individual is evaluated by testing it against a model of the underlying problem, and the quality is

5 Dr R.E.J. Keller, Robert.E.Keller@gmx.net, Leiden University, LIACS, Leiden Institute Advanced Computer Sciences, Leiden, The Netherlands

represented as a numerical value that is attached to the individual.
Recombination generates an offspring individual from more than one parental individual by combining partial information from the parents into the complete information of the offspring individual. A 'lucky' recombination may result in an individual that has a higher quality than any of its parents. Mutation copies a parental individual and then introduces a change into the copy which results in an offspring individual. A mutation represents a 'jump' within the search space over some distance, and a 'lucky' jump may hit an offspring individual that has a higher quality than its parents.
The described processes are stochastic which makes an evolutionary algorithm stochastic itself. For instance, mutation may perform a purely random change of the parental individual. For real-world problems, an evolutionary algorithm is usually guided by both deterministic and random heuristics. For example, problem-specific knowledge, e.g. experience or physical laws, may be represented as a deterministic heuristics.
As an evolutionary algorithm is a stochastic algorithm, there is no guarantee that it locates an acceptable solution, but, during its run time, an evolutionary algorithm often does exactly this. The average quality of the population increases and an acceptable local or even global optimum may be found.
Major classes of evolutionary algorithms are genetic algorithms, evolution strategies, genetic programming, evolutionary programming, and a synopsis may be found in [Baeck, 1997].

Genetic programming is emphasized here as it is increasingly being used for data mining. Genetic programming [Koza, 1992; Banzhaf, 1997] is a label for that set of evolutionary algorithms which, for the purpose of quality evaluation, represent a generated individual as an algorithm. This algorithm is then mapped onto an executable program, and genetic programming calculates the quality of the individual from the behavior this program displays, when fed with problem data. Individuals are created from a symbol set that contains operands and operators. With this in mind, genetic programming follows the basic scheme that is given in the Inset in the previous article.

A TOY PROBLEM: GENETIC PROGRAMMING FOR THE MODELING OF BLACK BOX BEHAVIOR

A toy problem will illustrate the practical use of genetic programming in a context that is also relevant to data mining. In particular, some typical classes of decisions will be introduced that are highly critical to the performance of the algorithm. For a deep discussion of the practical application of genetic programming technology to real-world problems see [Keller, 2001].
Let the problem be the identification of a relationship between the input and output data of a black-box system. The relationship is given as the function

$f(a,m,v,q)=a*a$, while, in the real world, such a — significantly more complicated — relationship would be unknown. Instead, we would expect the algorithm to identify the relationship, when being presented with system I/O data.
Three further parameters m,v,q have been introduced as noise, and all parameter values shall be real-valued and come from [0,1]. Due to the resulting real-valued five-dimensional parameter space, a vector of problem-representing data needed for evaluating the quality of an individual consists of four real input values and one real face output value. This output value represents the output as it is expected from a perfect individual, so that, for instance, an input vector (1.3,2.4,1.03,3) is paired with the output value 1.69, as $1.3^2=1.69$.
During quality evaluation of each individual, an input vector is fed to the individual, and the resulting output is compared to the face output value belonging to the input vector. An approximate solution, e.g. $a*a-m$, produces an error that is detected by this comparison. This step is repeated for all problem-representing vectors from a user-given training set that represents the underlying problem to the genetic programming algorithm. The errors an individual produces on all vectors from this set are input to a formula, e.g. the sum of the squared errors, which delivers a value that the algorithm interprets as the quality of the individual.
For the given simple problem, we define a population size of 50 individuals, and the run shall terminate after, say, 30 generations. As a rule of thumb it can be said that the harder a problem is, the higher the number of generated individuals has to be in order to locate an acceptable solution.

Symbol set

In the real world, the identification of a problem-relevant symbol set represents a major challenge to the user. The set must contain all symbols needed by the algorithm for generating an acceptable solution, and it should not contain irrelevant symbols, since these symbols increase the search space. In practice, the design of an advantageous symbol set is often impossible, because the decision maker does not fully understand the problem. Especially, in data mining, a real-world problem involves this classification of relevant, i.e. 'meaningful', versus irrelevant, i.e. 'noisy', data. Sophisticated genetic programming algorithms, e.g. [Keller, 2001], feature specific mechanisms for learning this classification in parallel to discovering an interesting relationship. For the toy problem, let us give the symbol set $\{m,v,q,a,+,*,-,/\}$ to the algorithm, which certainly allows for composing the perfect individual '$a*a$'.

Individual maximum size

Another important parameter to a genetic programming algorithm is the maximal size of a generated individual. This size must not be too small with respect to the representation of an acceptable solution, and it should not be too big in

order not to blow up the search space. For the toy problem, genotype size 7 is chosen, so that maximal-size individuals such as *a+m–q*q* or *a*a+m–m* may be generated by the algorithm.

Operator settings

Eventually, probabilities for the execution of genetic operators must be defined. For our simple problem, we may give the algorithm a rather exploratory nature by setting the probability of mutation being the next executed operator to 0.4, leaving 0.6 probability for reproduction. Real-world problems require a significantly more conservative search strategy with variation probabilities set to very low values. Sophisticated evolutionary algorithms often manage to come up with a heuristic dynamic adaptation of operator execution probabilities during a run.

Runs on the toy problem would typically produce perfect solutions, resembling *a*a*, *a*q/q*a* or *v–v+a*a*. Implicitly, neutral subterms such as the above *q/q* indicate to the decision maker that the involved parameters represent data that is irrelevant in the context of the underlying problem.

Evolutionary algorithms for data mining

Over the past six years, evolutionary algorithms have been increasingly recognized as flexible and powerful tools for approaching real-world data mining. An evolved structure with a high fitness represents a good model of the problem environment, which means that this model represents knowledge about the problem. In this sense, artificial evolution is an instance of knowledge discovery that delivers the information on the problem domain as a temporal sequence of increasingly knowledgeable entities. There is a wealth of structural representations used by evolutionary algorithms, especially since the introduction of genetic programming to evolutionary algorithms. Thus, classic representations, like decision trees generated by methods of rule induction, can be evolved as well as other representations for specific problem classes.

Conventional rule induction methods often show a troublesome tendency to act like hill climbers[6] in the multimodal fitness landscapes given by the high-dimensional data of data mining problems.

In other words, they apply a greedy approach to problems that rather call for a sophisticated combination of local search — that is exploitation — and global search, that is, exploration. Powerful instances of evolutionary algorithms provide exploitation and exploration of the search space and maintain a dynamic balance of these two major search phenomena over the run time of the algorithm.

Furthermore, an evolutionary algorithm, by its very nature an open-ended search process, can go on improving found solutions until being interrupted by the user. Opposed to this, classic data mining procedures often feature a certain

6 Hill climbing is a technique to discover peaks or optimum values, but has a danger of 'getting stuck' in a local optimum, See Section 6.2.14.

finite run time by design, thus leaving potential for solution improvement untapped.

As general problem solvers in arbitrary complex environments, off-the-shelf evolutionary algorithms also have properties that are detrimental to data mining. A prominent problem is their extensive run time on very large data banks, because such cases require the use of large populations which in turn draw heavily on computing resources during fitness evaluation. Another characteristic source of trouble is the difficulty of introducing domain knowledge to such standard evolutionary algorithms, e.g. the representation of side constraints is sometimes not straightforward. Thus, there is ongoing research into addressing these problems by extending evolutionary algorithms into tools for special problem domains and, more important, into fully self-adaptive tools.

As data mining is a field that is comprised of a great number of different problem domains and techniques, equally diverse tasks have to be met in the area of evolutionary algorithms. For instance, some major problem domains being tackled by evolutionary algorithms are clustering, classification, dependence modeling, symbolic regression, unstructured data mining, and time series analysis. As for the employed techniques, there are the use of distributed evolutionary algorithms, evolving software agents, and such various hybrids as fuzzy-evolutionary and neuro-evolutionary, to mention a few. As long as there are no fully autonomous evolutionary methods, the problem-specific manual tuning of parameters, such as representations, genetic operators and fitness functions, remains a tedious task.

The typical view of evolutionary algorithms today is still that of stand-alone systems that the user feeds with input, hoping for adequate output within an acceptable time frame, and being well aware of the nature of this tool. It is the conviction of the author that evolutionary algorithms will share the fate of electricity, microchips and the Internet in that they and their successors will become invisible ubiquitous powerful tools. With respect to data mining, for instance, there is ongoing research into integrating evolutionary algorithms into database systems, e.g. in order to use these methods as operators in complex queries.

Another aspect of evolutionary algorithms which makes them interesting data mining tools is their implicit handling of uncertainty, which is due to the continuously scalable selection probabilities of individuals. As with fuzzy technologies, there is no need for clean input from the user, because evolutionary algorithms can react to noisy or missing data with results that are more or less 'good'.

The function of an evolutionary algorithm in the entire process of data mining[7], i.e. data preprocessing (selection, pruning, etc.), knowledge discovery, postprocessing of discovered knowledge, is not fixed. There are several examples for the use of evolutionary algorithms for all of these stages in different applica-

7 See Chapter 6.1.

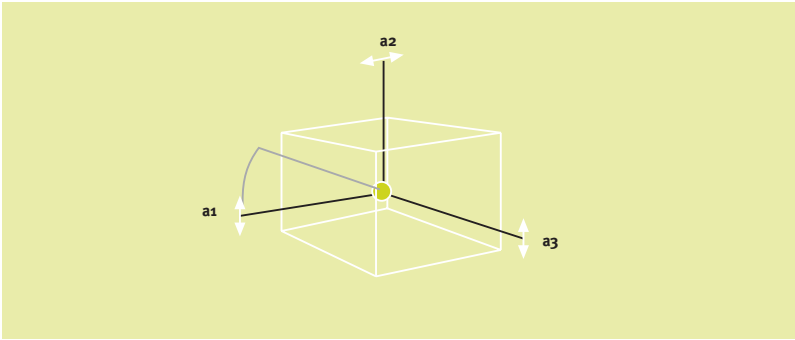
tions, so that an evolutionary algorithm can be freely combined with other, e.g. classic data mining methods.
Finally, there will be a brief examination of a slow auto-catalytic process with great impact for both process partners. Recently, evolutionary algorithms have been recognized as tools for the mining of the exponentially growing very large biotechnological data bases in order to perform, for instance, genome analysis. This will enhance the understanding of processes essential to the evolution and ontogeny of organic life forms, which in turn will provide the field of evolutionary computation with new metaphors.

In the context of data mining, genetic programming is an especially interesting variant of evolutionary algorithms due to the representational diversity and the algorithmic structures of the individuals. For instance, decision trees are classic data mining structures, and general tree structures are typical instances of genetic programming individuals. Often, closed, that is, analytical, expressions as they are delivered by regression methods cannot capture the true nature of a data mining domain. For instance, iterative or even recursive processes, like stock-market agent behavior, can be most naturally modeled by Turing-complete genetic programming individuals. Please note: data mining means so much more than viewing the world in those worn-out If-Then rules.
The representational and structural diversity employed by genetic programming simplifies the human interpretation of the extracted knowledge significantly. Particularly patterns of interest — for instance, of temporal or spatial nature — may be represented in abstract, e.g. symbolic form.
Using genetic programming instead of another evolutionary algorithm is an elegant approach to the problem at hand, because the result not only represents discovered knowledge but, in unchanged form, the application of this knowledge. This is because the result, by definition of genetic programming, represents a model of the mined data, which has been fed as training input to the individuals as well as to a computer program, whose execution applies the knowledge. The following example section shall illustrate this situation in the context of a control problem.

A real-world problem: genetic programming for the modeling of spacecraft-attitude maneuvers

Howley describes a genetic programming application from the domain of data mining and optimal control [Howley, 1996] where a general pattern in data representing spacecraft-attitude maneuvers is to be discovered. This resulting pattern can then be interpreted as a control law governing the performance of the considered class of maneuvers. As a result in genetic programming represents a computer program, this program may be used as a controller to perform an attitude maneuver, i.e. a craft reorientation, in minimal time. This means that, given

Figure 1
Spacecraft-attitude problem.



an initial craft attitude and a desired final attitude, the program must rotate the craft in minimal time by turning it around its three pair wise orthogonal body axes.

For each axis, there is an actuator, i.e. a craft component, which generates a positive or a negative torque that lets the craft rotate correspondingly around the axis. Each actuator is assumed to have a 'bang-bang' characteristic, so that it generates either a maximal positive or maximal negative torque.
The problem is relevant in practice, e.g. for satellite-based data transmission and observation. Typically, a satellite has to keep an optical system in a more or less exact focus on a planetary target area. Thus, if the satellite is moving relative to the planetary surface, it must reorient permanently or in sufficiently short-time intervals in order to stay focused.
The application concentrates on two maneuver types, i.e. rest-to-rest and rate-limited non-zero (RLNZ) terminal velocity, where rate means velocity. A rest-to-rest maneuver begins and ends with zero angular rates, that is, there is no rotation about any axis before or after the maneuver. Often, however, a RLNZ maneuver is needed i.e. before or after the maneuver, the craft is rotating about one or more axes with certain angular rates. For instance, when a sensor on the satellite surface is supposed to track a moving ground-based system, RLNZ maneuvers are required. The maximal angular rates are limited by the maximal forces generated by the actuators. In particular, a rest-to-rest maneuver represents a RLNZ maneuver with zero initial and final rate.
According to a theorem of Euler, a solid object can get from an arbitrary attitude into another arbitrary attitude by a rotation through a certain angle about a corresponding axis called the eigenaxis. Let us consider the non-trivial case of the eigenaxis not being identical to a body axis. Instead of performing a rotation sequence about several body axes, the craft may just rotate through a certain angle about the corresponding eigenaxis by operating one or more actuators in parallel. In summary, a certain attitude maneuver requires identifying the corresponding eigenaxis, the angle and direction for turning the axis, and the final turning rate.

If you have trouble imagining this situation, take a rubber eraser and hold it in some initial attitude, then move it into some final attitude by a sequence of body axis rotations. For any initial and final attitude, you can stick a pin through the eraser such that, when rotating the pin appropriately, the eraser will turn from the initial into to the final attitude: the pin represents the eigenaxis.

Since an actuator has a bang-bang characteristic, a maneuver consists of a sequence of actuator-switching commands. Each command switches one or more actuators to a positive or negative torque respectively. For instance, a command may switch actuators one and two to negative and three to positive. Thus, a command can be represented by a torque vector (t_1, t_2, t_3) with each component designating a positive or negative torque. This vector is the output of a control law that takes as input the desired final eigenaxis and rate and the current eigenaxis and rate.

For the application of a control law at the start of the maneuver, the initial eigenaxis and rate are the current parameters to the corresponding controller. The controller becomes active for the first time, computes and applies a torque vector, and the corresponding actuator activities result in a new current eigenaxis and rate. The controller becomes active again, and so on, until, in the case of a good controller, the desired final eigenaxis and rate are reached which terminates the control loop.

The control problem for rest-to-rest maneuvers has a known numerical solution, but the computation of this solution takes some time, whereas the problem must be solved in real time. There is no sense in computing a close to optimal solution that, when finally available, cannot be applied since it is out of date. Thus, an approximate but real-time solution is required, and it is realized by the control loop which makes the craft move incrementally into the final eigenaxis and rate.

In the context of genetic programming, an individual is a control law, so that genetic programming delivers expressions that are used as control law within the control loop. The symbol set contains variables for the described input and output parameters, the operators +,–,×, and a protected division operator that behaves graciously on division by zero. It also contains sign inversion, the absolute-value function, an if-a-less-b-then-action-else-action construct, and three ADF (automatically defined function) symbols. Genetic programming may use such a symbol as a label for an algorithm it has created, which it may then reuse by employing the label as an operand of an individual.

The training data consists of the initial eigenaxis and rate as well as the final eigenaxis and rate. Quality evaluation considered an individual as good, if the individual's application by the control loop resulted in the final eigenaxis and rate within user-given error bounds and before a user-given time-out. For rest-to-rest maneuvers, genetic programming runs went over 51 generations and used a population size of 5,000 individuals. For RLNZ maneuvers, the corre-

sponding values were 74 and 10,000, while, as genetic operators, 90% recombination and 10% copying were employed.
The approach produced a best result for the rest-to-rest maneuver within plus or minus 2% of the numerical solution. Additionally, this solution was generalized to solve further randomly generated maneuvers. As for the RLNZ maneuver, genetic programming produced a best solution that solved all test cases.

Applying evolutionary algorithms to data mining tasks

The amount of information on applying evolutionary computation to data mining has grown exponentially over the last four years. Thus, the reader is strongly recommended to place corresponding database queries to http://www.ira.uka.de/bibliography, the collection of computer science bibliographies, or one of its numerous international mirror sites. The data mining core tasks mentioned in Chapter 3.1 of this book can all be combined with evolutionary algorithms. Queries like 'evolutionary AND < task name >' will return a wealth of information on the respective subject.

References

- Baeck, T., D.B. Fogel, Z. Michalewicz. (1997). Handbook of Evolutionary Computation. IOP Publishing and Oxford University Press. Bristol. Evolutionary Programming 6.08. 2.59. hps. da
- Banzhaf, W., P. Nordin, R.E. Keller, F.D. Francone. (1998). Genetic Programming — An Introduction; On the Automatic Evolution of Computer Programs and its Applications. Morgan Kaufmann. Dpunkt Verlag
- Freitas, A. (1997). A Genetic Programming Framework for Two Data Mining Tasks: Classification and Generalized Rule Induction. In: J.R. Koza, K. Deb, M. Dorigo, D.B. Fogel, M. Garzon, H. Iba, R.L. Riolo. (1997). Genetic Programming 1997: Proceedings of the Second Annual Conference. pp96-101. Stanford University, San Francisco, Morgan Kaufmann, San Francisco
- Freitas, (1999). A. Data Mining with Evolutionary Algorithms: Research Directions. AAAI Press, Orlando, Florida
- Howley, B. (1996). Genetic Programming of Near-Minimum-Time Spacecraft Attitude Maneuvers. In: J.R. Koza, D.E. Goldberg, D.B. Fogel, R.L. Riolo. Genetic Programming 1996: Proceedings of the First Annual Conference. pp98-106. Stanford University, MIT Press, Cambridge MA
- Keller, R.E., W. Banzhaf. The Evolution of Genetic Code on a Hard Problem. Proceedings of the Genetic and Evolutionary Computation Conference. GECC-2001, San Francisco
- Koza, J.R. (1992). Genetic Programming: On the Programming of Computers by Natural Selection. MIT Press, Cambridge MA
- Ryu, T-W., Ch.F. Eick. (1996). Deriving Queries From Examples Using Genetic

Programming. In: E. Simoudis, J. Wei Han, U. Fayyad. The Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. http://www.cs.uh.edu/twryu/papers/kdd96.ps. http://www.aaai.org:80/Press/Proceedings/KDD/1996/kdd-96.html

- Teller, A., M. Veloso. (1995). Program Evolution for Data Mining. Sushil Louis. JAI Press. The International Journal of Expert Systems. pp216-236. http://www.cs.cmu.edu/afs/cs/usr/astro/public/papers/Astro-ESJ.ps
- Wong, M.L., K. Sak Leung. (2000). Data Mining Using Grammar Based Genetic Programming and Applications. Kluwer Academic Publishers. http://www.wkap.nl/book.htm

6.2.16 Fuzzy logic techniques

Robert Babuška[1]

Introduction

Most data mining methods rely on rather standard modeling techniques developed in the statistics, computer science, artificial intelligence and engineering communities. These include non-linear regression models, decision trees, neural networks, fuzzy logic and a number of other techniques. Fuzzy logic methods are among the tools that were introduced relatively recently and are gradually gaining more attention. Fuzzy models describe systems by establishing relations between the variables by means of if-then rules, such as 'If valve is wide open then pressure is low.' The ambiguous linguistic terms in the logical predicates like 'wide open' or 'low' are defined with the help of fuzzy sets, i.e. sets with non-sharp and overlapping boundaries.

Traditionally, fuzzy systems were built solely based on expert knowledge in a linguistic form and were used in expert systems or knowledge-based controllers [Mamdani, 1975; Driankov, 1993]. Recently, we have witnessed an increasing interest in the construction of fuzzy systems from data and applying them in areas like data mining, pattern recognition, systems identification or modeling [Bormans, 1997; Babuška, 1998; Hellendoorn, 1997]. In such applications, fuzzy systems can serve as an alternative or can complement other inductive methods, like neural networks, machine learning or statistical techniques. The most prominent feature that distinguishes fuzzy systems from black-box methods, such as neural networks, is their transparency and interpretability. Fuzzy models are ideally suited for explaining solutions to users, especially in situations where they do not have a strong mathematical background.

The purpose of this chapter is to give a concise introduction into fuzzy logic modeling. The basic notions of fuzzy sets, fuzzy propositions, logic connectives and their use in if–then rules are explained. Fuzzy clustering is briefly mentioned as a technique commonly used to induce fuzzy models from data and an application example from literature is described. Suggestions for further reading are given at the end of the chapter.

Fuzzy sets and systems

For illustrative purposes, consider a simplified fuzzy model to predict algae growth in a lake ecosystem. This model was induced from a data set containing observations from nine different shallow lakes around the country [Sanchez, 1996; Setnes, 2000]. The model consist of if–then rules in the following form:

1 Dr R. Babuška, r.babuska@et.tudelft.nl, Control Engineering Laboratory, Faculty of Information Technology and Systems, Delft University of Technology, Delft, The Netherlands, http://Lcewww.et.tudelft.nl/~babuska

Formula 1

If temperature (*t*) is *High* **and** nitrogen concentration (*N*) **not** *Low*
then algae growth is *Fast*

The if-part of the rule (called the antecedent) specifies under which conditions for the temperature (*t*) and the nitrogen concentration (*N*) the rule holds. The then-part of the rule (called the consequent) defines the output of the model under the given condition. Both the antecedent and the consequent contain linguistic terms such as Low, High and Fast. These terms represent in an approximate, quantized way the magnitude of the different variables involved. Fuzzy sets are used to define the meaning of the linguistic terms. This type of fuzzy model is called the linguistic or Mamdani fuzzy model [Zadeh, 1973; Mamdani, 1977].
Another type of fuzzy system is the Takagi–Sugeno (TS) fuzzy model [Takagi, 1985], in which the consequent is not a fuzzy set, but rather a local (usually linear) regression model:

Formula 2

If *t* is *High* **and** *N* **not** *Low*
Then algae growth = $a \cdot t + b \cdot N + c$

where *a*, *b* and *c* are parameters, whose values will be different for the different rules. In this way, complex non-linear functions can be approximated by a collection of multiple liner submodels.

Fuzzy sets

Consider the variable 'temperature' in the above example. The domain of this variable can be partitioned, for instance, into three sets, denoted by linguistic terms Low, Medium and High. By using conventional sets, a temperature measurement can be classified as High, if it is greater than a certain threshold, say 10°C. Mathematically, this set is represented by its characteristic function $\mu_{High}(t)$:

Formula 3

$$\mu_{High}(t) = \begin{cases} 1 & t \geq 10 \\ 0 & otherwise \end{cases} .$$

This characteristic function is depicted in Figure 1.

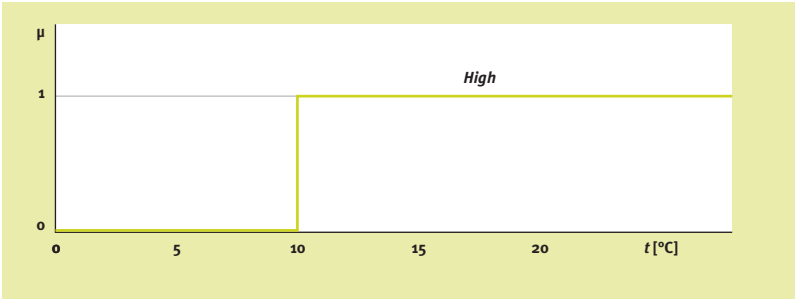


Figure 1
Characteristic function of the set of 'high temperature'.

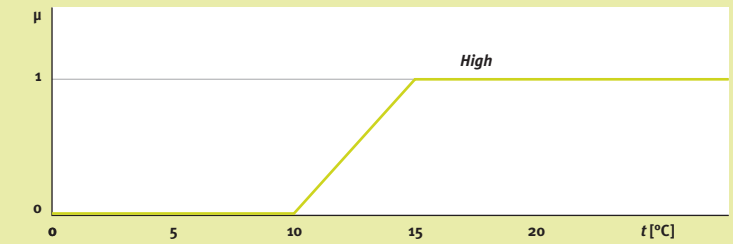
In this setting, a particular temperature either fully belongs to the set or it does not. For example, 10.01˚C is classified as a high temperature, while 9.95˚C is not. This abrupt transition from membership to non-membership is usually rather inconvenient in the modeling of real-world systems. It is much more reasonable to accept a gradual transition from non-high temperatures, to high temperatures. This concept can be represented by a fuzzy set [Zadeh, 1965], which is obtained by generalizing the characteristic function of conventional sets. The characteristic function of a fuzzy set is called the membership function and it ranges over the closed interval [0,1].

Formula 4

$$\mu_{High}(t) = \begin{cases} = 1 & t \text{ is a full member of the set} \\ \in (0,1) & t \text{ is a partial member of the set} \\ = 0 & t \text{ is not a member of the set} \end{cases}$$

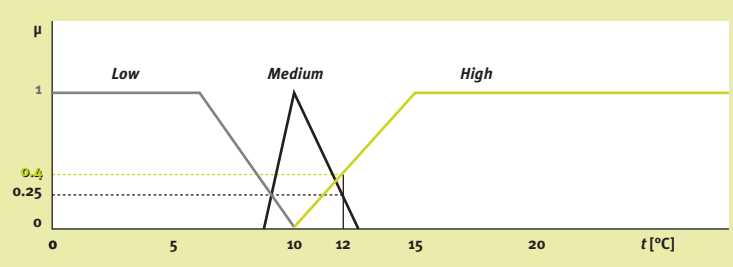
A possible membership function for 'high temperatures' is depicted in Figure 2.

Figure 2
Membership function of the fuzzy set 'high temperature'.



According to this definition — temperatures below 10˚C definitely do not belong to the set of high temperatures (membership degree zero), between 10˚C and 15˚C — there is a gradual transition from non-membership to full-membership, while temperatures above 15˚C definitely belong to the set. Because of this gradual transition, the temperatures can be partitioned into overlapping intervals as shown in Figure 3.

Figure 3
Partitioning of the temperature domain into three fuzzy sets.



It is thus possible that a particular point in the domain belongs to more than one set with various degrees of membership. For instance, $t = 12$˚C belongs to

the set of high temperatures with membership 0.4 and to the set of medium temperatures with membership 0.25. As fuzzy sets are used to denote linguistic terms, the variable 'temperature' is also called a linguistic variable.
The membership degree is at the same time the truth value of the corresponding fuzzy logic proposition, in our example, 'temperature is high.' The truth value of combined propositions involving logic connectives ('and', 'or', 'not') is computed by using fuzzy logic operators discussed in the next paragraph.
In a computer implementation, fuzzy sets are usually represented in two ways: on discrete[2] domains as a list of ordered pairs

$$\left\{\left(x_i, \mu\left(x_i\right)\right) \middle| \forall x_i \in X\right\}$$

and on continuous domains through an analytical formula for the membership function, e.g.

$$\mu(x) = 1/\left(1 + x^2\right), \forall x \in \Re$$

Figure 4 shows some commonly used types of membership functions. The singleton set is a degenerated fuzzy set: the fuzzy-set representation of a non-fuzzy (crisp) number. The position and the shape of the membership functions depend on the particular application at hand.

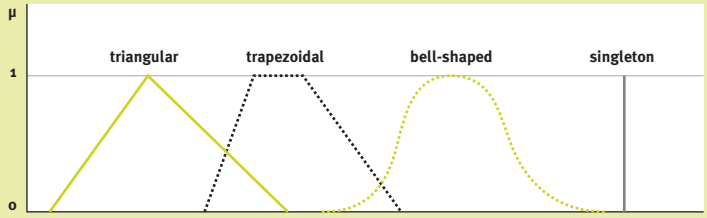


Figure 4
Various types of membership functions.

Fuzzy set and fuzzy logic operations

In order to work with fuzzy propositions and rules, such as in Formula 1 and 2, operators for the logic connectives 'and', 'or', 'not' and the 'if–then' implication are needed. To this end, operators from conventional set theory and Boolean logic have been extended to their fuzzy equivalents. For instance, the fuzzy 'and' operator (conjunction, intersection of fuzzy sets) is computed as the minimum of the corresponding membership degrees and the 'not' operator (negation, complement) is one minus the membership degree. In our example, the truth value β of the antecedent of Formula 1 is shown in Formula 5, indicating that β will take on the value of whichever of the two values within the brackets is lowest.

2 With stepwise values for x, like 1,2,3,4.

Formula 5

$$\beta = \min\left(\mu_{High}(t), 1 - \mu_{Low}(N)\right) \quad .$$

Operators are available for all common logic operations like the disjunction, implication, etc. Moreover, some special operators have been introduced in fuzzy sets and fuzzy logic. Examples are linguistic hedges to represent expressions like 'very high temperature', or operations with fuzzy relations.
On the basis of logic operators and implications, a fuzzy inference mechanism implements the reasoning with fuzzy if–then rules for some given input (or output) data. Most fuzzy inference methods are basically interpolation algorithms for which it has been proven that fuzzy systems can approximate smooth functions to any degree of accuracy [Wang, 1992]. In this sense, the use of fuzzy systems in data-driven modeling and approximation is theoretically justified, although it should be noted that fuzzy systems are less effective approximators than multi-layer neural networks (i.e. they need more parameters for a comparable accuracy). The main advantage of fuzzy systems, however, is the transparency of the knowledge representation.

Fuzzy clustering

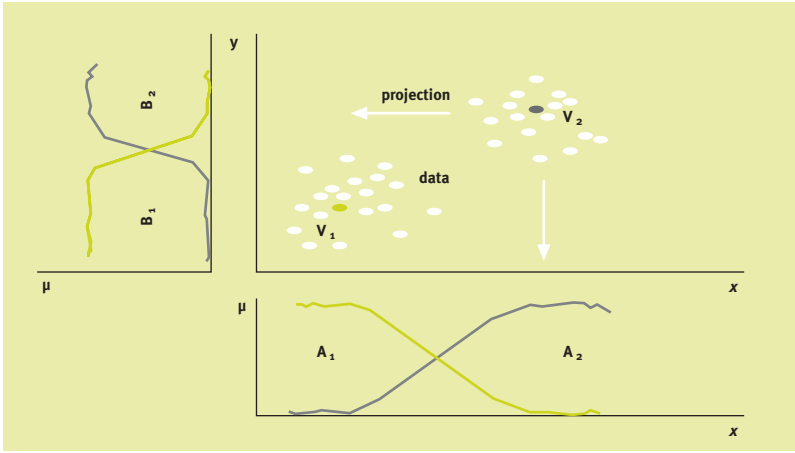
In pattern recognition and data mining applications, the concept of fuzzy membership can be used to represent the degree to which a given data sample is similar to another sample or some prototype. Based on this similarity, data points can be clustered into groups (clusters). The potential of fuzzy clustering algorithms to reveal the underlying structures in data in an unsupervised manner has been successfully exploited in a wide variety of applications ranging from image processing [Bezdekl, 1992] to customer segmentation [Setnes, 2001].
The local structures in data discovered by fuzzy clustering can be conveniently represented by means of fuzzy if-then rules. Figure 5 shows an illustrative example with two clusters found in the two-dimensional product space $X \times Y$. After clustering, each data point is assigned a pair of membership degrees that tell how much that point belongs to each cluster. The membership degrees are organized in the fuzzy partition matrix:

Formula 6

$$U = \begin{bmatrix} \mu_{11} & \mu_{12} & \dots & \mu_{1n} \\ \mu_{21} & \mu_{22} & \dots & \mu_{2n} \end{bmatrix} \quad .$$

The clusters are characterized by their prototypes v_1 and v_2, which are also obtained by clustering. There are many different clustering methods, the most well-known is the fuzzy c-mean algorithm [Bezdek, 1981].
In order to interpret the (multidimensional) data, rules can be extracted from the fuzzy partition matrix. Each cluster induces one if–then rule of the type:

Figure 5
Extraction of linguistic rules by fuzzy clustering.



If *x* is *A*
Then *y* is *B*

where the membership functions A_i (B_i) are obtained by projecting the partition matrix on the *x* (*y*) axis, see Figure 5.

The model builder or the user can then assign suitable linguistic labels to the sets A_i and B_i in order to have an interpretable model. In our simple example, this can be, for instance:

If *x* is Small
Then *y* is Small

If *x* is Large
Then *y* is Large

Models of the TS type can be constructed in a similar way, for instance by combining fuzzy clustering and weighted least-square estimation (Babuska, 1998).

Application example

An application of fuzzy clustering and decision making approaches to data mining for customer selection in direct marketing is briefly described. Details can be found in (Setnes, 2001). The data set was obtained by a large Dutch financial services provider from a direct-marketing campaign for a savings product. The goal was to obtain a customer preference model that could be used in future mailing campaigns to select customers who would be most likely to respond to the mailing. The data set used to train the model contained 16.525 client records, each described by 170 features and a known response indication (9.1% clients responded). The features include indicators like the possession of cer-

tain types of accounts, credit cards, account balances, age category, etc. Fuzzy clustering and decision-making techniques were used both to preprocess the data (imputation of missing values, reduction of the number of features) and to develop a model to predict the client response. The algorithms selected a subset of 28 features ordered according to their importance for the prediction. Finally, three most important features (total savings balance and the balance on two particular types of accounts) were used to compare the achieved results with the CHAID decision tree model, which represents the state of the art method used in practice. The gain chart in Figure 6 shows that fuzzy modeling outperforms the standard approach. It is interesting to note that good results were also obtained by combining the two techniques: fuzzy data preprocessing and feature selection and CHAID decision tree modeling.

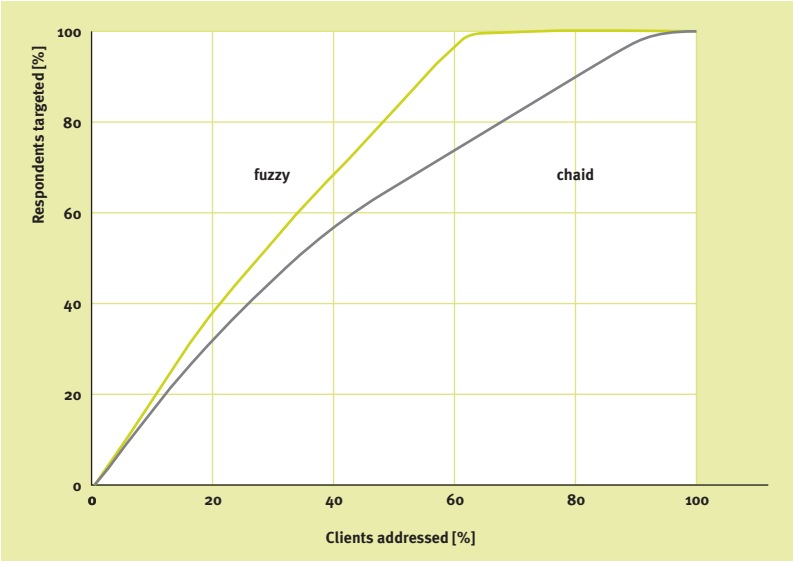


Figure 6
Comparison of the response gain charts obtained by fuzzy modeling and the CHAID method.

A rule base can be extracted from the parametric model obtained by clustering. This allows the inspection, validation and possibly also extension or adaptation of the model by a human expert. The rules are of the form:

R_i : **If** feature$_j$ is A_{ij} **Then** response = c_i

where A_{ij} are linguistic labels defined by membership functions that were induced from the fuzzy partition in a way similar to Figure 5. The consequent parameters c_i are numbers between 0 and 1 indicating the possibility that clients with the given characteristic will respond. The rules and the corresponding membership functions can be found in [Setnes, 2001].

Conclusions and further reading

Techniques based on fuzzy set theory and fuzzy logic are promising tools for data mining applications. As opposed to predominantly data-driven black-box methods, like artificial neural networks, fuzzy logic techniques are to a certain degree transparent and thus support the interactive participation of the user. However, rather than regarding them as competitors to other techniques, one should realize that the different methods can be effectively combined and can complement each other. No single tool or set of tools should be used exclusively. For any given problem, the nature of the data will affect the choice of tool. Consequently, one needs a variety of tools and technologies in order to find the best possible model. Numerous practical examples of successful synergies between different modeling paradigms can be found in the current literature. This chapter hopefully fulfilled its aim to serve as a basic reference and an 'appetizer' for the interested reader. More detailed and comprehensive treatment of the different subjects related to fuzzy sets and fuzzy logic can be found on the CD-rom shipped with this book, in many textbooks and research monographs, as well as scientific journals.

A number of popular-scientific and practically oriented books are available for those who do not want to jump right into the mathematics [Kosko, 1993; McNeill, 1992; McNeill, 1994; Cox, 1994]. For a comprehensive mathematical treatment of fuzzy set theory, the following classical monographs can be recommended: [Dubois, 1980; Klir, 1988; Zimmermann, 1996; Klir, 1995]. Readers interested in a deeper treatment of fuzzy clustering may refer to the books by [Bezdek, 1981] and [Jain, 1988]. A more recent overview of different clustering algorithms can be found in [Bezdek, 1992]. Various learning algorithms and models based on the combination of fuzzy and neural techniques are presented in [Brown, 1994; Jang, 1997; Hellendoorn, 1997]. Data mining applications of fuzzy and other computational intelligence techniques are given by [Maimon, 2000; Kandel, 2001]. Journals regularly publishing scientific papers in the area of fuzzy set techniques include IEEE Transactions on Fuzzy Systems, IEEE Transactions on Systems, Man and Cybernetics, Fuzzy Sets and Systems, among many others.

References

- Babuška, R. (1998). Fuzzy Modeling for Control. Kluwer Academic Publishers, Boston
- Bezdek, J.C. (1981). Pattern Recognition with Fuzzy Objective Function. Plenum Press, New York
- Bezdek, J.C., S.K. Pal. (eds.). (1992). Fuzzy Models for Pattern Recognition. IEEE Press, New York
- Bormans, N.W., M. Setnes, U. Kaymak, H.R. van Nauta Lemke. (1997). Application of Fuzzy Sets to Finance and Insurance. Proceedings of

Interfaces'97. Montpellier, France. pp189–191
- Brown, M., C. Harris. (1994). Neurofuzzy Adaptive Modelling and Control. Prentice Hall, New York
- Cox, E. (1994). The Fuzzy Systems Handbook: A Practitioner's Guide to Building, Using, and Maintaining Fuzzy Systems. Academic Press, Boston, MA
- Driankov, D., H. Hellendoorn, M. Reinfrank. (1993). An Introduction to Fuzzy Control. Springer Verlag, Berlin
- Dubois, D., H. Prade. (1980). Fuzzy Sets and Systems: Theory and Applications. Mathematics in Science and Engineering **144**. Academic Press
- Hellendoorn, H., D. Driankov. (eds.). (1997). Fuzzy Model Identification: Selected Approaches. Springer Verlag, Berlin
- Jain, A.K., R.C. Dubes. (1988). Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs
- Jang, J.-S.R., C.-T. Sun, E. Mizutani. (1997). Neuro-Fuzzy and Soft Computing; a Computational Approach to Learning and Machine Intelligence. Prentice-Hall, Upper Sadle River
- Kandel, A., M. Last, H. Bunke. (eds.). (2001). Data Mining and Computational Intelligence **68**. Physica-Verlag, Studies in Fuzziness and Soft Computing
- Klir, G.J., T.A. Folger. (1988). Fuzzy Sets, Uncertainty and Information. Prentice-Hall, New Jersey
- Klir, G.J., B. Yuan. (1995). Fuzzy Sets and Fuzzy Logic; Theory and Applications. Prentice Hall
- Kosko, B. (1993). Fuzzy Thinking: The New Science of Fuzzy Logic. Warner
- Maimon, O., M. Last. (2000). Knowledge Discovery and Data Mining – The Info-Fuzzy Network (IFN) Methodology. Kluwer Academic Publishers, Boston
- Mamdani, E.H., S. Assilian. (1975). An Experiment in Linguistic Synthesis with a Fuzzy Logic Controller. International Journal of Man-Machine Studies **7**:1–13
- Mamdani, E.H. (1977). Application of Fuzzy Logic to Approximate Reasoning Using Linguistic Systems. Fuzzy Sets and Systems **26**:1182–1191
- McNeill, D., P. Freiberger. (1992). Fuzzy Logic: The Discovery of a Revolutionary Computer Technology. Simon and Schuster
- McNeill, F., M. Thro, E. Thro. (1994). Fuzzy Logic: A Practical Approach. Academic Press, Boston
- Sanchez, D., H.F.P. van den Boogaard. (1996). Applications of the Sugeno–Takagi Fuzzy Model. Project Report X189. Delft Hydraulics. Delft, The Netherlands
- Setnes, M., R. Babuška. (2000). Transparent Fuzzy Modeling. A. Kent and J.G.
- Setnes, M., U. Kaymak. (2001). Fuzzy Modeling of Client Preference from Large Data Sets: an Application to Target Selection in Direct Marketing. IEEE Transactions on Fuzzy Systems **9** (1):153–163

- Takagi, T., M. Sugeno. (1985). Fuzzy Identification of Systems and its Application to Modeling and Control. IEEE Transactions on Systems, Man and Cybernetics **15** (1):116–132
- Wang, L.-X. (1992). Fuzzy Systems are Universal Approximators. Proceedings IEEE International Conference on Fuzzy Systems 1992. San Diego, USA. pp1163–1170
- Williams. (eds.). Encyclopedia of Computer Science and Technology **43** (Supplement 28):pp303–335. Marcel Dekker, New York
- Zadeh, L.A. (1965). Fuzzy sets. Information and Control **8**:338–353
- Zadeh, L.A. (1973). Outline of a New Approach to the Analysis of Complex Systems and Decision Processes. IEEE Transactions on Systems, Man and Cybernetics **1**:28–44
- Zimmermann, H.-J. (1996). Fuzzy Set Theory and its Application. Kluwer Academic Publishers, Boston

6.2.17 Rough sets

Helge G. Solheim[1]

The Rough Set theory has become a popular theory in the field of data mining and classification. The theory was introduced by Zdzislaw Pawlak in the early 1980s, and based on this theory one can propose a formal framework for the automated transformation of data into knowledge. Pawlak has shown that the principles for learning by examples can be formulated in the basis of his theory. An important result of the theory is that it simplifies the search for dominating attributes leading to specific properties, and may therefore be used for classification of new objects.

The rough set theory is mathematically relatively simple. Despite this, it has shown its fruitfulness in a variety of data mining areas. Among these are information retrieval, decision support, machine learning, and knowledge based systems. A wide range of applications utilize the ideas of the theory. Medical data analysis, aircraft pilot performance evaluation, image processing, and voice recognition are a few examples.

Almost inevitably the database used for data mining will contain imperfections, such as noise, unknown values or errors due to inaccurate measuring equipment. The rough set theory comes handy for dealing with these types of problems, as it is a tool for handling vagueness and uncertainty inherent to decision situations. An advantage of the rough sets methodology over the Bayesian approach is that no assumptions about the independence of the attributes are necessary, neither is any background knowledge about the data.

In this section, a set of definitions from the world of rough sets is given. An example is shown in parallel with most definitions. Part of the section is from [Aasheim, 1996], and some of the definitions from [Mollestad, 1995].

Information system

We will look at an information system containing a set of objects. Each object has a number of attributes with attribute values related to it. The attributes are the same for all objects, but the attribute values may differ. An information system is thus more or less the same as a relational database.

Information System

An Information System (IS) is an ordered pair $A = (U,A)$ where U is a nonempty finite set of objects — the Universe, and A is a nonempty, finite set of elements called Attributes. The elements of the Universe will further be referred to as Objects. Every attribute $a \in A$ is a total function $a:U \rightarrow V_a$, where V_a is the set of allowed values for the attribute (its range).

A Decision System is an IS $A = (U,A)$ for which the attributes in A are further

1 H.G. Solheim, hgs@computas.no, Computas AS, Lysaker, Norway, www.computas.no

classified into disjoint sets of condition attributes C and decision attributes D. ($A = C \cup D$, $C \cap D = \varnothing$).

Decision system — deciding on income

An example of a decision system, is shown in Table 1. As one might expect, it is a two-dimensional data table. The rows represent objects, while the columns represent attribute values belonging to these objects.

Table 1
Example of an information system with 5 objects with some attributes.

	Studies	Education	Speaks French	Income
1	no	good	yes	high
2	no	good	yes	high
3	yes	good	yes	none
4	no	poor	no	low
5	no	poor	no	medium

In this information system there are 5 persons (objects) with attributes reflecting each person's situation in life. Assume the intention is to discover rules predicting what degree of income a person has, depending on a few attributes describing him. The attribute income is therefore selected as a decision attribute (or dependent attribute). The rest of the attributes, studies, education, and speaks French are the condition attributes (independent attributes). This situation with only one decision attribute is by far the most common, and if there is more than one decision attribute, they may often be found one at a time.

Discerning objects

The next definition introduces the concept of an indiscernibility relation. If such a relation exists between two objects, it means that all their attribute values are identical with respect to the attributes under consideration, and thus cannot be discerned between (distinguished) by using the considered attributes.

Indiscernibility relation

With every subset of attributes $B \subseteq A$ in the IS $A = (U,A)$, an equivalence relation IND(B) is associated, called an Indiscernibility relation, which is defined as follows:

$$\mathrm{IND}(B) = \{(x,y) \in U^2 \mid a(x) = a(y) \text{ for every } a \in B\}$$

By $U/\mathrm{IND}(B)$ is meant the set of all equivalence classes in the relation IND(B).

For the decision system given earlier, a calculation of $U/\mathrm{IND}(C)$ gives the following result:

$U/\mathrm{IND}(\{studies, education, speaks\ French\}) = \{\{1, 2\}, \{3\}, \{4, 5\}\}$

One can see that the objects are grouped together, and that the groups consist of objects that cannot be distinguished, when using the selected set of attributes. With a (equivalence) class is meant such a group (or set). The classes in tabular form are shown in Table 2. Class E_1 comes from objects 1 and 2, class E_2 object 3, while class E_3 comes from objects 4 and 5. Note that E_3 has two objects with different decision attribute values.

Table 2
The example divided in equivalence classes.

	Studies	**Education**	**Speaks French**
$E1$	no	good	yes
$E2$	yes	good	yes
$E3$	no	poor	no

Discernibility matrix

A discernibility matrix is a matrix with one row and column for each equivalence class. In the matrix, the condition attributes that can be used to discern between the classes in the corresponding row and column are inserted.

Discernibility matrix

For a set of attributes $B \subseteq A$ in $A = (U,A)$, the discernibility matrix $M_D(B) =$

$\{m_D(i,j)\}_{n \times n}$, $1 \leq i,j \leq n = \ |\ U/\mathrm{IND}(B)\ |$, where
$m_D(i,j) = \{a \in B \mid a(E_i) \neq a(E_j)\}$ for i,j = 1,2,...,n

The entry $m_D(i,j)$ in the discernibility matrix is the set of attributes from B that discern object classes $E_i, E_j \in U/\mathrm{IND}(B)$.

From the previous example, one can observe that the only attribute with a different value for classes E_1 and E_2 is studies. This attribute is therefore placed in the corresponding places in the matrix. Naturally, the matrix will be symmetric due to the fact that the attributes that differ in value for objects a and b, differ 'the other way around' in value for b and a. Completing the calculation of the discernibility matrix results in the matrix shown in Table 3.

Table 3
The discernibility matrix of the example.

	E1	***E2***	***E3***	
$E1$	∅			
$E2$	studies	∅		
$E3$	education, speaks French	studies, education	speaks French	∅

If some of the classes have the same decision value, one might decide not to discern between these classes. By doing so, attributes are not added to the matrix for classes with the same decision value. This can result in more simplistic rules, if any classes have the same decision value. In the example IS presented earlier this is not an option, since all classes have different decision values.

Discernibility functions

From the discernibility matrix, two useful functions may be calculated. These are called discernibility functions and relative discernibility functions and will be used for finding the most important attributes in the information system.

Discernibility Function

The Discernibility Function $f(B)$ of a set of attributes $B \subseteq A$ is

$$f(B) = \bigwedge_{i,j \in \{1 \ldots n\}} \bigvee \bar{m}_D(E_i, E_j)$$

where $n = |U/\mathrm{IND}(B)|$, and $\bigvee \bar{m}_D(E_i,E_j)$ is the disjunction taken over the set of Boolean variables $\bar{m}_D(i,j)$ corresponding to the discernibility matrix element $m_D(i,j)$.
For example, this is:

$f(C) = studies \wedge (education \vee speaks\ French) \wedge$

$(studies \vee education \vee speaks\ French)$

Definition of the Relative Discernibility Function

The Relative Discernibility Function $f(E, B)$ of an object class E, attributes $B \subseteq A$ is

$$f(E,B) = \bigwedge_{j \in \{1 \ldots n\}} \bigvee \bar{m}_D(E, E_j)$$

where $n = |U/\mathrm{IND}(B)|$.
This implies that the discernibility function $f(B)$ computes the minimal sets of attributes required to discern any equivalence class from all the others.
Similarly, the relative discernibility function $f(E, B)$ computes the minimal sets of attributes required to discern a given class E from the others.

For the example above, the following relative discernibility functions can be calculated:

$f(E_1, C) = studies \wedge (education \vee speaks\ French)$
$f(E_2, C) = studies \wedge (studies \vee education \vee speaks\ French)$
$f(E_3, C) = (education \vee speaks\ French) \wedge (studies \vee education \vee speaks\ French)$

Dispensability

An attribute a is said to be dispensable or superfluous in $B \subseteq A$ if $\mathrm{IND}(B) = \mathrm{IND}(B–\{a\})$, otherwise the attribute is indispensable in B. If all attributes $a \in B$ are indispensable in B, then B is called orthogonal.

For example, over the set of classes the attribute values for the attributes education and speaks French go hand in hand. Whenever education is good, speaks French is yes, and whenever education is poor, speaks French is no. Thus, $\mathrm{IND}(C) = \mathrm{IND}(C–\{\textit{speaks French}\}) = \mathrm{IND}(C–\{\textit{education}\})$. The only indispensable attribute in our example is studies.

Reducing representation

The data in the information system can only be used to discern classes to a certain degree. However, not all attributes may be required in order to be able to do so. The next definitions, which are based on the discernibility functions, help in finding useful combinations of attributes and may later be used to generate classification rules.

Reducts and Relative Reducts

A Reduct of B is a set of attributes $B' \subseteq B$ such that all attributes $a \in B–B'$ are dispensable, and $\mathrm{IND}(B') = \mathrm{IND}(B)$. The term $\mathrm{RED}(B)$ is used to denote the family of reducts of B. The set of prime implicants of the discernibility function $f(B)$ determines the reducts of B.

The set of prime implicants of the relative discernibility function $f(E, B)$ determines the Relative Reduct of B. The term $\mathrm{RED}(E,B)$ denotes the family of relative reducts of B for an object class E.

What this implies is that a relative reduct contains enough information to discern objects in one class from all the other classes in the information system. A reduct contains enough information to discern all classes from all the others.

To find the reducts for our example, the discernibility functions are employed. Each function is minimized to a sum of products form, as shown below.

$$
\begin{aligned}
f(C) &= \textit{studies} \wedge (\textit{education} \vee \textit{speaks French}) \wedge \\
&\quad (\textit{studies} \vee \textit{education} \vee \textit{speaks French}) \\
&= (\textit{studies} \wedge \textit{education}) \vee (\textit{studies} \wedge \textit{speaks French})
\end{aligned}
$$

$$
\begin{aligned}
f(E_1, C) &= \textit{studies} \wedge (\textit{education} \vee \textit{speaks French}) \\
&= (\textit{studies} \wedge \textit{education}) \vee (\textit{studies} \wedge \textit{speaks French})
\end{aligned}
$$

$$
\begin{aligned}
f(E_2, C) &= \textit{studies} \wedge (\textit{studies} \vee \textit{education} \vee \textit{speaks French}) \\
&= \textit{studies}
\end{aligned}
$$

$$f(E_3, C) = (education \vee speaks\ French) \wedge (studies \vee education \vee speaks\ French)$$
$$= education \wedge speaks\ French$$

This gives the desired relative reducts. For instance, RED(E_1, C) = {{studies, education}, {studies, speaks French}}.

The relative reducts are minimal, because each discernibility function was minimized. A minimal (relative) reduct is thus a reduct in which none of the attributes may be removed without removing the reduct property.

Upper and Lower Approximation

The next definition given is fundamental to the concept of rough sets, since it addresses the central point of the approach, the vague classes. These are the ones with more than one value for the decision attribute.

Lower and Upper Approximation

The Lower Approximation $\underline{B}X$ and the Upper Approximation $[(\overline{B})]X$ of a set of objects $X \subseteq U$ with reference to a set of attributes $B \subseteq A$ (defining an equivalence relation on U) may be defined in terms of the classes in the equivalence relation, as follows:

$$\bar{B}X = \bigcup \left\{ E \in U/\mathbf{IND}(B) \middle| E \cap X \neq \varnothing \right\}$$

$$\underline{B}X = \bigcup \left\{ E \in U/\mathbf{IND}(B) \middle| E \subseteq X \right\}$$

called the B-upper and the B-lower approximation of X, respectively. The region $BN_B(X) = [(\overline{B})]X - \underline{B}X$ is called the B-boundary (region) of X.

The lower approximation of X is the collection of objects which can be classified with full certainty as members of the set X, using the attribute set B. Similarly, the upper approximation of X is the collection of objects that may possibly be classified as members of the set X. The boundary region comprises the objects that cannot be classified with certainty to be neither inside X, nor outside X, again using the attribute set B. Properties of these approximations are given in [Pawlak, 1991].

From reducts to rules

Rules represent dependencies in the dataset, and represent extracted knowledge, which can be used when classifying new objects not present in the original information system. When the reducts were found, the job of creating definite rules for the value of the decision attribute of the information system was

practically done. To transform a reduct (relative or not) into a rule, one only has to bind the condition attribute values of the object class from which the reduct originated to the corresponding attributes of the reduct. Then, to complete the rule, a decision part comprising the resulting part of the rule is added. This is done in the same way as for the condition attributes.

The rules in our example are as follows.

E_1: studies=no ∧ education=good → income=high
studies=no ∧ speaks French=yes → income=high
E_2: studies=yes → income=none
E_3: education=poor → income=?
speaks French=no → income=?

The 'rules' derived with basis in E_3 do not specify the resulting attribute value for income, since it is not the same for all the objects in the class. It may therefore be called a vague category. A better way of presenting this than through a question mark would be to say e.g. that if education is poor, then there is a 50% chance that income is low, and that there is a 50% chance that income is medium.

If a new object is introduced to the example data set with the decision value missing, one could attempt to determine this value by using the previously generated rules. If exactly one rule which fits is found, the classification is straightforward. This also implies that the object is in the lower approximation of the class to which it is classified as belonging to. For objects contained in the boundary region of different classes, no such consistent decision can be made. Then the rough membership function described next may be used.

Rough Membership Function

The Rough Membership Function (RMS) expresses how strongly an element x belongs to the rough set X in view of information about the element expressed by the set of attributes B.

Rough Membership Function

For $A = (U, A)$, $x \in U$, $X \subseteq U$, attributes $B \subseteq A$, the Rough Membership Function for a class $E \in U/\mathrm{IND}(B)$ is

$$\mu_B(E, X) = \frac{|E \cap X|}{|E|}, 0 \leq \mu_B(E, X) \leq 1$$

Consider the information system in Table 4.

Table 4
Example information system.

	a	*b*	*d*	# of objects
$E_{1,1}$	0	0	0	99
$E_{1,2}$	0	0	1	1

The class E_1 is split into two decisions, but one decision value is dominating. Calculation of the RMS for the object in class $E_{1,2}$ gives the value 0.01. For the other class, it is 0.99. When the set X is a decision class and the objects in U are indiscernible, the RMS can be used as a measure of accuracy, or validity.

Rough sets for classification

In the previous chapter, we discussed how to create a decision support system for model selection and parameter setting. This was described as a classification problem, and we wanted a tool to use it. In addition, we wanted the classification to be based on old cases.

Rough sets can very well be used for this purpose with the definitions of this chapter. Classification vectors can be stored for old cases in a so-called decision system. Based on this, reducts can be generated, and decision rules created. These rules can then easily be used to classify a new classification vector with the decision missing (i.e. which model to use or what parameter values.). Reducts, and thus rules, are calculated quite fast for modest to large sized tables, if the number of attributes is not too large. Therefore, this calculation can be done quite often. As the old cases change, the decision system changes as well, and a recalculation of the rules will make the rules adapt to the new cases.

The conclusion of this section is that rough sets is a well suited tool for classification.

This article is taken from [Solheim, 1996].

References

- Aasheim, O.T., H.G. Solheim. (1996). Rough Sets as a Framework for Data Mining. Norwegian University of Science and Technology, Trondheim, Norway. Project Report
- Mollestad, T. (1995). Learning Propositional Default Rules Using the Rough Set Approach. In: A. Aamodt, J. Komorowski. (eds.). Scandinavian Conference on Artificial Intelligence. IOS Press
- Pawlak, Z. (1991). Rough Sets. Kluwer Academic Publishers
- Solheim, H.G. (1996). Decision Aids for Model Selection and Parameter Setting in Operations Management. Norwegian University of Science and Technology, Trondheim, Norway. Master Thesis

6.2.18 Support vector machines

Maarten van Someren[1]

Support vector machines (SVM's) are a family of methods that combine several principles. The most popular form is for classification of objects described as numerical vectors. The problem is to find a plane that maximally separates two classes on the basis of vectors that are labeled positive or negative. The standard statistical method for this is (linear) discriminant analysis (see Section 6.2.3) which constructs a plane that maximizes the separation between two classes. In technical terms, the standard method maximizes the distance between means of the two classes (when these are projected on the plane). Other popular approaches to this problem are neural networks of various flavors trained by minimizing misclassifications (Section 6.2.8).

SVM's follow a different approach, characterized by two goals:

- to construct a class boundary that minimizes both the expressiveness of the function (expressed as the VC dimension[2]) and the expected error on the training data. Expressiveness (or 'capacity') means the range of patterns that can be expressed by the class of boundary functions that are considered. SVM is thus not limited to planes (linear boundaries), but can also be applied with non-linear shaped boundaries.
- to find the class boundary that maximizes the 'margin': the distance between the data points of different classes that are closest to the boundary between the classes. The data points that are on the 'right side' of the boundary and that are closest to the boundary are called 'support vectors'.

The functions that are used to construct the boundary are called the 'kernel' functions. The SVM approach can be extended to class boundaries other than linear functions of the original dimensions. In this case, the method outlined above is followed, using the VC dimension of the kernel to try increasingly expressive classes of functions. The computations for finding the boundary will become more complex. Consider the example given in Figure 1.

Suppose also that our kernel function is simply the inner product of vectors v and w: $v.w = v_1w_1 + v_1w_1$.

SVM now finds a line that separates positive and negative data points such that the margin, the distance between the line and the data points closest to the line but in the 'right' part of the space is maximized. Data points that end up on the line are called the support vectors. Actually, SVM maximizes the margin itself without constructing the boundary line (or in more dimensions, the hyper plane) that separates the classes. The line can be reconstructed from the support vectors.

1 Dr M. van Someren, maarten@swi.psy.uva.nl, The Universiteit van Amsterdam, Faculty of Social and Behavioural Sciences, Department of Psychology, Amsterdam, The Netherlands

2 VC dimension (Vapnik Chernonenkis dimension): a measure for the complexity of a learning task. Given a domain, a set of possible instances I and a hypothesis set H such that each hypothesis classifies the instances in I, the VC dimension of the combination of I and H is the size of the largest subset S of I that is 'shattered' by H. 'Shattered' means that for each possible way of dividing S into classes, there exists a corresponding hypothesis in H. This may seem a strange concept, because the size of the hypothesis space and of the instance space seem more direct measures for the complexity, but the VC dimension is very useful for spaces with an infinite number of instances and/or hypotheses.

Figure 1
Example distribution of points in two dimensions.

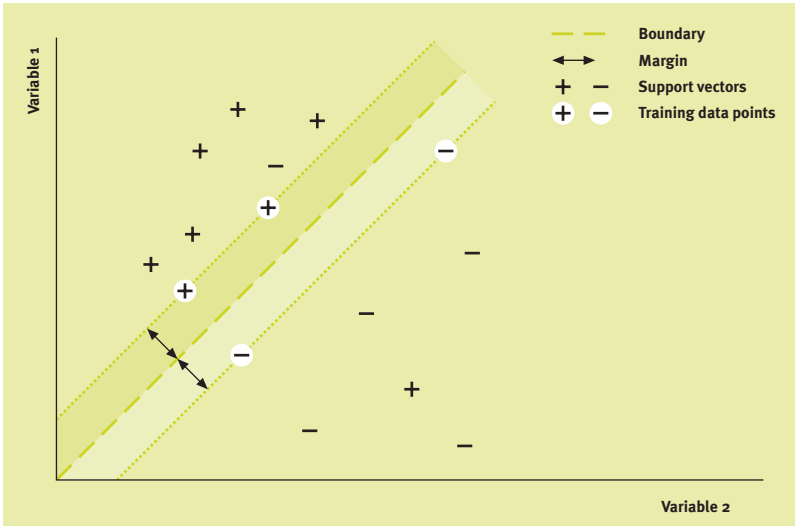
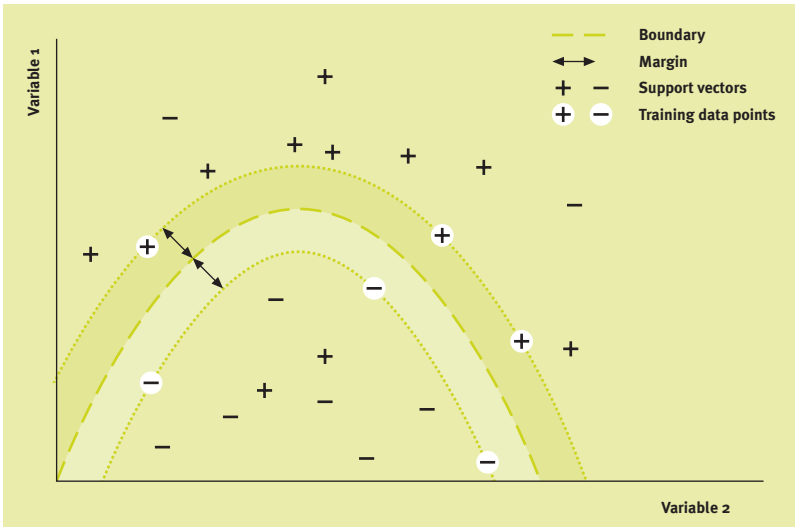


Figure 2
Curved boundaries from a non-linear kernel function.



Suppose that we use a kernel function that gives non-linear projection. In terms of the example above, this allows margins that are not linear but have a different shape.

The method has several strengths: it can be applied in combination with different kernel functions, it is computationally much cheaper than comparable methods (like discriminant analysis and log-linear regression), because it uses only data points close to the boundary and it is possible to see the relation between a model that was constructed and the data on which the model is based. SVM's have proven their functionality in applications with large amounts of data, such as text recognition which often involves large numbers of training data with many features.

Applications

SVM can be used for regression, classification, and density estimation problems. Popular applications can be found for problems with many features and many training data, such as text classification and image recognition. Some examples are [Joachims, 1998] where SVM was applied to classification of (web) documents.

Computation

Computing the support vectors involves selecting a kernel (the function for computing the margin, effectively the shape of the boundary) and computing the margin and the support vectors. Algorithms for finding the support vectors can exploit the fact that only a relatively small subset of all data points needs to be considered. SVM algorithms use this to allow a wide range of possible boundary shapes. The third idea is that the data points themselves are not needed, but only their relative positions or even only the relative positions of vectors that are close to possible boundaries. SVM uses the inner product of vectors to represent the relative positions. In fact, a kernel function does not map dimensions to new dimensions, but (pairs of) data vectors to inner products. This can reduce computation of the model at the price of increased costs for applying the model to new examples, because the method does not output the actual boundary. Some well-known kernels are: polynomial classifiers, radial basis function (RBF) classifiers and two layer sigmoid neural nets.

Tools and information

Implementations of SVM and more information are available from the Web, for example at http://www.kernel-machines.org/.

References

- Burges, C.J.C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery **2**:121-167
- Cristianini, N., J. Shawe-Taylor. (2000). An Introduction to Support Vector Machines, (and Other Kernel-Based Learning Methods). Cambridge University Press
- Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: C. Nedellec, C. Rouveirol. (eds.). Proceedings of the 10th European Conference on Machine Learning. pp137-142, Springer Verlag
- Vapnik, V.N. (1995). The Nature of Statistical Learning Theory. Springer Verlag

6.2.19 Combining classifiers: voting, stacking, bagging and boosting

Wojtek Kowalczyk[1]

Introduction

As we have seen in the previous sections there are many methods for solving classification problems. Some of these methods (e.g. neural networks) are non-deterministic and for different runs they may produce different models. Even a deterministic method may produce several models when applied to several samples of the training data. One may also apply various methods (e.g. decision trees or naive Bayes) to the same data set, ending up with multiple models. No matter how the models are generated one may ask a question: should we choose the best model or is it perhaps better to combine all the models into one that outperforms all its components? And how could such a combination be constructed? These questions, raised about 20 years ago, led to the discovery of a number of techniques and algorithms that substantially improved the accuracy of 'classical' classification algorithms. Also our understanding of the notorious bias-variance error decomposition has deepened considerably.

In this paper we will present, very informally, the main methods of combining classifiers: voting, stacking, bagging and boosting. We will pay special attention to boosting. This method, essentially different from the former three, constructs a sequence of classifiers in such a way that the n-th classifier tries to 'correct' mistakes made by its (n-1) predecessors. Using this idea one may 'boost' the performance of any, even very weak, base classifier to an arbitrarily height. It can be shown that if the base classifier performs just slightly better than random choice, then it may be boosted to reach accuracy 100% (on the training set), see [Freund, 1997]. Although this result says nothing about the accuracy of the boosted classifier on the test data, numerous experiments have demonstrated its superiority over any 'single classifier' approach [Ditterich, 2000; Bouwer, 1999; Opitz, 1999]. Boosting algorithms have been also successfully applied to a number of challenging classification problems in real-life challenging classification problems: fraud detection with credit card transactions [Fan, 1999], text mining [Schapire, 2000], optical character recognition [Schwenk, 1997].

In 1996 Leo Breiman, a prominent researcher in the fields of machine learning and statistics, expressed his enthusiasm about AdaBoost (a specific variant of boosting) by proclaiming it to be 'the best off-the-shelf classifier in the world'. Today, 5 years later, boosting algorithms are still considered to be most powerful, although their position is threatened by (much more complicated and numerically less stable) Support Vector Machines (see Section 6.2.18).

1 Dr W. Kowalczyk, w.kowalczyk.cs.vu.nl, Faculty of Mathematics, Department of Artificial Intelligence, The Vrije Universiteit Amsterdam, The Netherlands, http://www.cs.vu.nl/~wojtek/

There is a huge body of literature on combining classifiers available on the internet. For example, the site www.boosting.org contains 169 references to papers on boosting and related topics (December 2001). Most of them are directly available in electronic form. A very concise introduction to boosting can be found in [Freund, 1999; Friedman, 2000] contains mathematical analysis of several variants of boosting algorithms. Finally, [Ditterich, 2000; Bouwer, 1999; Opitz, 1999] present results of numerous benchmarking experiments.

Voting

Let us consider a binary classification problem and let us suppose that we have developed a number of classifiers $f_1(x)$, $f_2(x)$, ..., $f_k(x)$. We will refer to these classifiers as base classifiers. Each of these classifiers is supposed to discriminate between 2 classes, say A and B, but obviously, each classifier can make some mistakes. The simplest way of combining these classifiers is to use the majority voting strategy: to determine a class of an instance x we apply all classifiers to it and count the number of resulting A's and B's. The most frequent class label is then assigned to x.

This simple scheme can be generalized by attaching some weights to each classifier, so 'more reliable' classifiers are weighted more highly. Very often, classifiers return a number between 0 and 1 rather than just a class label A or B. This number reflects the classifier's confidence in its prediction and can also be used for weighting votes. Thus, instead of just counting the number of votes for A and B we could now calculate the total confidence in A and B.

There are some theoretical results that demonstrate that these simple voting strategies result in classifiers that, under some conditions, are not worse than their components. When basic classifiers overfit the data, voting usually substantially improves the accuracy [Tresp, 2001].

Let us note that the majority voting scheme may in some situations fail. For example, if base classifiers correctly recognize non-overlapping fragments of the data, taking the 'maximum of votes' would be much better than taking the 'sum of votes'. More general, we may notice that the way in which votes should be combined can be viewed as a separate learning task. A general technique that is based on this observation is called stacking.

Stacking

As mentioned above, it is not always clear how to assign weights to different classifiers. It may strongly depend on the situation. Therefore, instead of making an 'educated guess' or running a number of trial-and-error experiments we may simply involve yet another learning algorithm that would learn how to com-

bine outputs from all classifiers to make the final decision. There are at least two questions to be answered first: What should be the form of this 'meta classifier'? What should be the training set?

Numerous experiments have demonstrated that the meta classifier should be a 'relatively global and smooth' classifier [Wolpert, 1992]. In practice a linear discriminant or a simple neural network will do the job. The second question turns out to be more difficult. A straightforward use of the whole training set usually leads to overfitting. Therefore, it is recommended to the application ofy a hold-out method is recommended, where a part of the data is used for training base classifiers and the remaining data is used for training the meta classifier. A more sophisticated method involves n-fold cross validation, where the data is split into a collection of n train and test sets, or, in case of very small data sets, leave-one-out cross validation. More details can be found in [Wolpert, 1992].

Bagging

Both voting and stacking can be used when relatively large data sets are available (so one can afford to taking multiple samples of the training sets), or when there is no risk of overfitting. However, in some cases we don't have much data and then the risk of overfitting is quite high. [Breiman, 1996] proposed the following solution to this problem: given a training set with N cases make K 'inexact copies' of it (e.g. K=100) and then develop K classifiers, one for each 'copy' of the training set. Next, combine all classifiers with help of majority voting. Each 'copy' is created by randomly drawing N cases from the training set, with replacement.

The whole procedure is called bagging (for 'bootstrap aggregation') and it performs very well in situations when the training set is relatively small and the base classifiers tend to overfit the data [Opitz, 1999]. Therefore, bagging can be viewed, in addition to cross validation and the bootstrap method [Efron, 1993], as yet another technique of avoiding overfitting. An extensive comparison of these methods can be found in [Kohavi, 1995].

Boosting

The three schemes described above start with a collection of classifiers that are developed independently from each other and then combine them into the final classification procedure. Boosting is based on a different idea. It works with a weighted training set, i.e. each case has a certain weight. Initially, all cases have the same weight, 1/N, where N is the number of cases in the training set. We start with building a single classifier $f_1(x)$ by applying our base learner to the training set. Then weights of all cases are modified according to predictions made by $f_1(x)$: weights of cases that are misclassified by $f_1(x)$ are increased,

weights of correctly classified cases are decreased. The next classifier, $f_2(x)$, is built by the same base learner and it tries to minimize the weighted error. In this way $f_2(x)$ 'pays more attention' to cases that are misclassified by $f_1(x)$. The procedure is repeated M times, where M is a parameter: in the n-th step all cases are re-weighted according to the outputs of $f_n(x)$ and a classifier $f_{n+1}(x)$ that minimizes the weighted error is found. Finally, all the classifiers are combined:

$$F(x) = \sum_{i=1}^{M} g_i f_i(x) \quad ,$$

where g_i's are constants that are calculated in a separate process.

It should be noticed that although most 'base learner' algorithms (e.g. decision tree inducers, naive Bayes, neural networks) were not designed to deal with weighted data, they can be easily generalized to handle such data. [Breiman, 1998] proposed another solution to this problem: instead of explicitly minimizing the weighted error we may draw (with replacement) N cases from our training set (of size N) using weights as a probability distribution, and then apply 'classical' learning algorithms to such a biased sample. He called this variant of boosting algorithm arcing. Taking into account that sampling costs extra time and introduces non-determinism to the whole process it is not surprising that arcing is less popular than boosting.

There are several details that should be filled in: What do we mean by 'weighted error'? How do we update weights? How do we compute weights of classifiers? Etc.

There are many possible answers to these questions and consequently, many variants of boosting algorithms have been developed: Discrete AdaBoost, Real AdaBoost, LogitBoost and Gentle AdaBoost, to mention the most representative ones (see [Friedman, 2000]). To give the reader some idea about possible choices we will describe one of these variants, the Gentle AdaBoost algorithm, in detail. In our opinion this algorithm is the most elegant one, numerically stable, easy to implement, and it outperformings all standard algorithms on certain classification tasks [Kowalczyk, 1999].

Gentle AdaBoost

Let us consider a training set $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$, where vectors x_i represent values of input variables and y_i class labels (-1 or +1). Let $w_1, w_2, \ldots, w_N$ denote weights that are assigned to training cases. We will always assume that all weights are non-negative and that they sum up to 1. Finally, let us suppose that we have a learning algorithm, 'base learner', that, when applied to a weighted training set, finds a classifier f(x) that minimizes the weighted squared error on the training set:

$$\sum_{i=1}^{N} w_i (f(x_i) - y_i)^2$$

To give a specific example of such a base learner, let us assume that $x_1, \ldots, x_N$ are numeric, i.e. $x_i \in R^n$, and that each variable can take a few values only: $v_1, v_2, \ldots, v_k$. Now let us consider a very simple classifier that is called a decision stamp. A decision stamp is a function of the form:

$$f(x) = \begin{cases} v_L, & if \ (x)_i < val \\ v_R, & otherwise \end{cases}$$

where $(x)_i$ denotes the i-th component of vector x.
In other words, a decision stamp simply checks if the i-th component of x is smaller than a certain value val and returns v_L or v_R, accordingly. Let us note that given i and val, the optimal values of v_L and v_R (in the least squares sense) can be easily calculated: v_L is just $\sum w_j y_j$, where the summation is taken over all x's with $(x)_i < val$; to calculate v_R the summation is taken over all remaining x's. Our base classifier will systematically search for the best decision stamp by calculating for all combinations of i, i=1, ..., n and v_j, j=1, ..., k the optimal values of v_L and v_R and the corresponding weighted error. A decision stamp that minimizes this error is returned as a result.
Now, we are ready to describe the Gentle AdaBoost algorithm [Friedman, 2000].

1 Initialize all weights to 1/N, F(x)=0.
2 Repeat for m=1,2, ..., M:

   a find best $f_m(x)$
   b update weights $w_i := w_i \exp(-y_i f_m(x_i))$ and renormalize them
   c $F(x) := F(x) + f_m(x)$

3 Output the classifier $\text{sign}(F(x)) = \text{sign}(\sum f_m(x))$.

Let us note that the weight update rule really does what it is supposed to: if $f_m(x_i)$ is consistent with y_i (i.e. $f_m(x_i)$ and y_i have the same sign) the product $y_i f_m(x_i)$ is positive and w_i is multiplied by a number smaller than 1; otherwise the product is negative and w_i is multiplied by a number that is bigger than 1. Moreover, let us note that all the intermediary classifiers $f_m(x)$ have the same weight, i.e. $g_i=1$, for i=1, ..., M.

It can be proven [Friedman, 2000] that in essence Gentle AdaBoost optimizes Shapire-Singer exponential loss function $\sum w_i \exp(-y_i f_m(x_i))$ using Newton stepping.

An illustrative example

To illustrate working of Gentle AdaBoost let us consider the task of separating all the points that are inside the circle given by the equation $x^2+y^2 < 0.7^2$ from points that are outside. Decision stamps have now the form of tests: 'is $x < a$?' or 'is $y < b$?' (see Figure 1). Thus our boosting algorithm is supposed to find a collection of lines that are parallel to x- or y- axis such that a linear combination of corresponding stamps would approximate our circle. This apparently difficult task is quickly accomplished by running the Gentle AdaBoost algorithm for 10000 iterations: the percentage of misclassified cases drops from 38% (after the first iteration) to about 4.5% (after the last iteration).

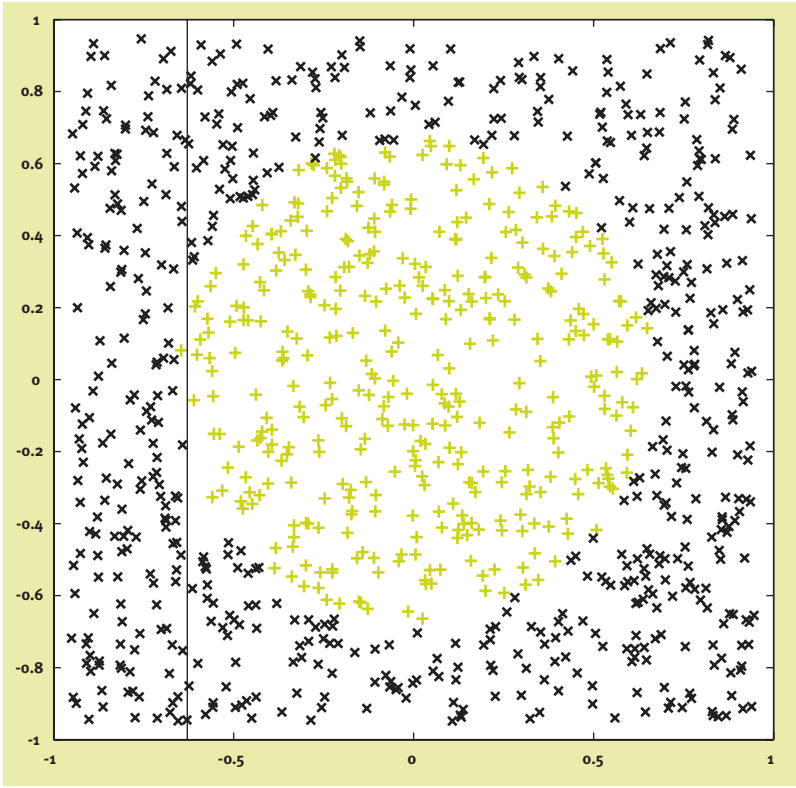


Figure 1
A classification problem: separate points that are inside the circle form points that are outside using a linear combination of simple decision stamps. An example decision stamp ($x < -0.62$) is also shown.

Let us take a look at our model F. It is just a linear combination of functions that involve either x or y. Let Fx and Fy denote the x-component and the y-component of F, respectively, see Figure 2.

As we can see, Fx and Fy correspond (modulo some constant factors) to x^2 and y^2 that were used in the definition of the circle! In other words, our algorithm has learned the (implicit) definition of the circle $Fx + Fy < 0$, just by looking at the training set from the x- and y-directions.

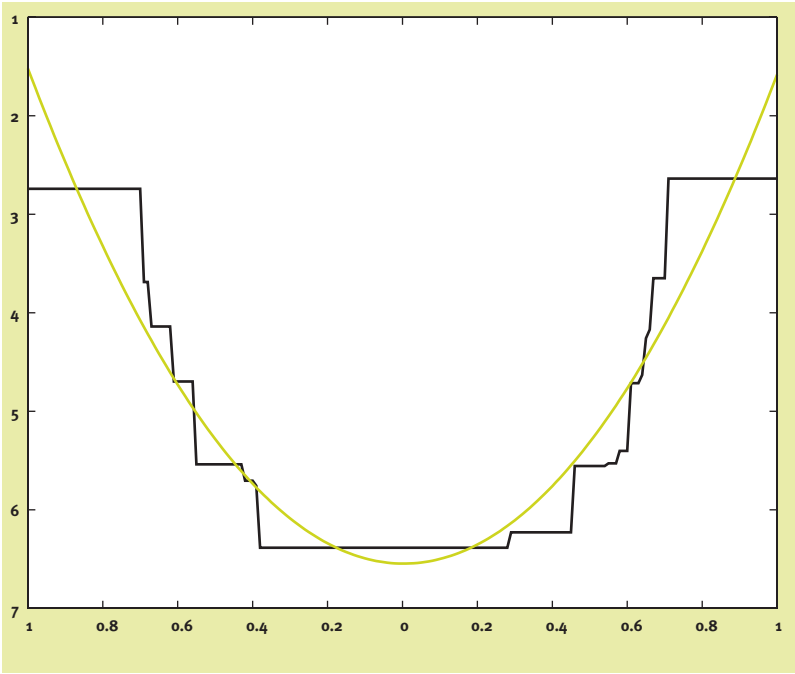


Figure 2
The Fx component of the final model (black line). For illustration purposes a least-square quadratic approximation is shown (green line). The Fy component is almost the same and is not displayed.

Conclusions

In this paper we have presented several techniques for combining classifiers. These techniques usually lead to classifiers with an increased accuracy. In their recent study Bauer and Kohavi [Bauer, 1999] report the average relative error reduction varying from 24% to 31% depending on the choice of base classifier. Similar results are reported by other authors, e.g. [Ditterich, 2000; Opitz, 1999]. In addition to increased accuracy boosting algorithms seem to be immune to overfitting: even after obtaining 100% accuracy on the training set further iteration of the algorithm does not influence the test error.

Although in this paper we have focused only on binary classification problems, all the algorithms have their multi-valued counterparts. The state of the art text classification algorithm BoosTexter [Shapire, 2000] can also handle multi-label classification problems, where a single case (a document) is allowed to belong to several classes. This algorithm can effectively handle data sets with thousands of attributes.

The only drawback of boosting algorithms that we are aware of is their time complexity. However, using the simple improvements proposed by [Friedman, 2000], one can speed up these algorithms by factor 10-50 without losing any accuracy.

References

- Bauer, E., R. Kohavi. (1999). An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. Machine Learning **36** (1/2):105-139
- Breiman, L. (1996). Bagging Predictors. Machine Learning **26** (2):123-140
- Breiman, L. (1998). Arcing Classifiers. The Annals of Statistics **26** (3):801-849
- Dietterich, T.G. (2000). An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. Machine Learning **40** (2):139-158
- Efron, B., R. Tibshirani. (1993). An Introduction to the Bootstrap. Chapman and Hall
- Fan, W., S.J. Stolfo, J. Zhang. (1999). The Application of AdaBoost for Distributed, Scalable and On-Line Learning. Unpublished Manuscript. http://www.boosting.org
- Freund, Y., R.E. Schapire. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences **55** (1):119-139
- Freund, Y., R.E. Schapire. (1999). A Short Introduction to Boosting. Journal of Japanese Society for Artificial Intelligence **14** (5):771-780. Appearing in Japanese. English Version. http://www.boosting.org
- Friedman, J., T. Hastie, R. Tibshirani. (2000). Additive Logistic Regression: a Statistical View of Boosting. The Annals of Statistics **38** (2):337-374
- Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Proceedings 14th International Joint Conference on Artificial Intelligence. Montreal. Morgan Kaufmann, 1137-1143
- Kowalczyk, W. (1999). Rough Data Modeling and Gentle AdaBoost: Two Approaches to the BENELEARN'99 Challenge. Winning Contribution. Unpublished. http://www.cs.vu.nl/~wojtek/
- Opitz, D., R. Maclin. (1999). Popular Ensemble Methods: An Empirical Study. Journal of AI Research **11**:169-198
- Schapire, R.E., Y. Singer. (2000). BoosTexter: A Boosting-Based System for Text Categorization. Machine Learning **39** (2/3):135-168
- Schwenk, H., Y. Bengio. (1997). Adaptive Boosting of Neural Networks for Character Recognition. Technical Report **1072**, Department d'Informatique et Recherche Operationelle, Université de Montreal. http://www.boosting.org
- Tresp, V. (2001). Committee Machines. In: Handbook on Neural Network Signal Processing. CRC Press
- Wolpert, D.H. (1992). Stacked Generalization. Neural Networks **5**: 241-259

6.2.20 Text mining techniques

Jeroen Meij, Antal van den Bosch[1]

Text mining is data mining on textual data; the extraction or discovery of knowledge from text. It is a young field, with relations to information retrieval, natural language processing, and data mining. Since text mining deals with natural language, it involves more than just 'letting the data speak for itself'. Natural language is actually an encryption of the information encapsulated in it; knowledge of the language (which is not always English) is needed for understanding, or decryption of text in that language. The decryption of bits of information from natural language is not straightforward. Not every word provides an equal amount of information; furthermore, the (partial) information encapsulated in a word is usually dependent on the surrounding text and on domain knowledge (up to world knowledge) of the observer. The computer-based equivalents of some of these human knowledge processes are syntactic and semantic disambiguation, lexicon and ontology augmentation, and discourse analysis [Hearst, 1999]. Thesauri and ontologies may be used to create a better correspondence between human understanding and computer understanding.

Text mining operations can be typically divided into four main steps:
- Preprocessing.
- Conversion to intermediary representations: decryption of the underlying information into a conceptual structure.
- Analysis: attaching relations among concepts and between concepts and documents.
- Knowledge representation.

Preprocessing

Text mining involves data preprocessing, preparing the textual data to allow for more precise results and faster processing. Some preprocessing steps:
- *Tokenization*: separating punctuation (periods, quotes, commas, etc.) from words, except from abbreviations; subsequently, identifying the boundaries between sentences in text.
- *Case folding*: removing differences between upper and lower case words, while retaining the meaningful capitalization of names.
- *Stemming*: reducing each word to its stem. This step is especially needed in the case of morphologically rich languages in which new word forms can be formed freely by compounding (Dutch, German), and languages with a rich inflectional morphology (Czech, Turkish).
- *Stop word removal*: removing stop words like can, will, do, the, etc., as far as they do not contribute to the meaning of the text.

1 Dr A van den Bosch, Antal.vdnBosch@kub.nl, ILK/Computational Linguistics, Tilburg University, The Netherlands, http://ilk.kub.nl/~antalb/

– *N-Grams*: instead of stemming or stop word removal, slices of N characters of words can be extracted, representing the original word by all the necessary slices.
– *Word-N-Grams*: slices of combinations of words ('United States', 'foreign affairs') can be identified as carrying more specific information than their parts.
– *Shallow parsing*: identifying syntactically functional units, such as noun phrases, named entities (proper names of persons, organizations, or locations) or phrases describing the time or location of action, and the assignment of subject and object roles to noun phrases with respect to main verbs.
– *Sectioning*: identifying meaningful sections in documents such as abstracts, concluding remarks, and references.

Intermediate representation

After preprocessing, documents are converted in intermediate representations, 'decrypting' the natural language of the documents. These intermediary representations can subsequently be processed through data mining algorithms. The most commonly used representation is the document vector representation.

Vectors

A document may be described by a vector of t terms, where each term corresponds to a word in the document, or to any other unit identified in preprocessing: N-Grams, Word-N-Grams, and syntactic phrases. Usually, a coefficient of significance is associated with the term, indicating its quantitative presence in the document, or its informational importance. Note that the vector size (dimensionality) equals the number of unique terms (words) present in the total collection. Some examples:

Term-frequency vector

Each document is represented by a vector, which can be seen as a vector in term space. Thus, a document might be represented by the term-frequency vector:

$$\vec{dtf} = \left(tf_1, tf_2, \ldots, tf_n\right)$$

where the frequency of each term in the document is multiplied by the term [Shankar, 2000]. The term frequency can be normalized to a value between 0 and 1 [Keen, 1991] resulting in a normalized term frequency or simplified to a (binary) Boolean weighting.

Inverse document frequency

The inverse document frequency (IDF) reduces the influence (weight) of a term appearing frequently in many documents. These frequent terms are considered to have less discriminating power. This is commonly done [Shankar, 2000] by

multiplying the frequency of each term *i* by , $\log\left(N/df_i\right)$ where N is the total

number of documents in the collection, and is the number of documents that contain the *i*th term (i.e. document frequency).

Combining these two leads to the tf-idf representation of the document, calculated as *tf x idf*. Tf-idf assigns high values to terms which are both important for describing the contents of a document and for discriminating between articles [Rauber, 1999; Salton, 1989].

$$\vec{d}_{tfidf} = \left(tf_1 \log\left(N/df_1\right), tf_2 \log\left(N/df_2\right), \ldots, tf_n \log\left(N/df_n\right)\right)$$

Augmented normalized term frequency

This is the term frequency divided by the maximum occurring *tf* and normalized to a value between 0.5 an 1 [Lucarella, 1988].

Information theory metrics

A counter-intuitive fact of natural language is that term frequency in itself is inversely proportional to term 'informativity'. The most frequent words in a language are the least informative, and vice versa. Where *tf x idf* attempts to counteract this through document frequencies, metrics exist that estimate the informativity of terms with respect to the desired outcome — the underlying meaning and conceptual structure of the text. Given a set of example texts of which the conceptual structure is analyzed (annotated) by experts, information theory [Shannon, 1948] offers good metrics for estimating informativity. We give two popular metrics: Information Gain, used frequently in other data mining areas and Odds Ratio [cf. Mladenic, 2001].

Information Gain

Information Gain is a quantification of the power of a term in decrypting meaning. The Information Gain of a term W is given as

$$InfoGain(W) = P(W)\sum_i P(C_i \mid W)\log\frac{P(C_i \mid W)}{P(C_i)} + P(\overline{W})\sum_i P(C_i \mid \overline{W})\log\frac{P(C_i \mid \overline{W})}{P(C_i)}$$

The outcome denotes W's contribution in bits (if the log has base 2) in decrypting the meaning of a sentence part, sentence, section, or document it is in, where C denotes the different symbolic classes (section types, document classes) in which meaning is expressed. P(W) is estimated by Ws observed frequency in the data, and P(C|W) is estimated by W's co-occurrence with class C. P($\overline{W}$) represents all terms except W.

Odds Ratio

Odds Ratio is a variant of Information Gain, which is particularly suited for two-class meanings, where one meaning is represented by a minority set of documents in a vast majority of irrelevant documents (a situation typical for the World Wide Web).

$$OddsRatio(W) = \log \frac{P(W \mid C_{pos})(1 - P(W \mid C_{neg}))}{(1 - P(W \mid C_{pos}))P(W \mid C_{neg})} \ .$$

Odds Ratio favors words that are good predictors of the minority class only.

Learning-based classification and information extraction

Term weighting metrics such as tf x idf and Information Gain are computed off-line on the basis of static document data bases. On-line estimation of these metrics on the basis of error during learning is also possible with machine learning algorithms; in document classification this has become the favored term-weighting method with the advent of more labeled example data bases. As with information-theory metrics, learning-based weighting techniques assume labeled data, and are therefore restricted to domains for which some predetermined labeling is available (document classes, keywords, named entities), which has been annotated in a document data base.

Examples of much-used machine learning algorithms for documents are support vector machines, Naïve Bayes, and the weighted k-Nearest Neighbor classifier [Yang, 1999; Joachims, 2002]. These algorithms are typically used for document classification, but are also used for information extraction tasks, such as the extraction and labeling of named entities (proper names of persons, organizations and locations) in text [Cucerzan, 1999].

The term weights developed through these learning-based methods thus reflect the importance of a term for a particular classification task, either at the document level (document classes, keywords, section types) or at the sentence level (named entities). All of these produced analyses: terms, weighted vectors of terms, their resulting classifications, and informationally salient entities in the text, can be joined to form enriched document representations for further analysis.

Analysis

Similarity of documents can be investigated by comparing the similarity of their informational representations. This comparison can be done using many different distance measures [Rhagavan, 1986]. Knowledge can be discovered through the investigation of found distances in textual databases, by using visualization or by browsing salient features in which documents are similar.

Two techniques often used for analyzing collections of documents are clustering and association discovery.

Clustering

Through clustering, documents are grouped together on the basis of their content. Clustering, if performed dynamically, is able to reflect changes in the document data base in the clustering results.
Commonly used clustering techniques for document clustering include hierarchical clustering (see Section 6.2.6), K-means [Steinbach, 2000], and two-dimensional clustering in Self-Organizing Maps [Kohonen, 1989].
Combined with sectioning, subtopical regions can be identified using distance measures [Mather, 2000].

Clustering can occur on data bases of unlabeled documents, but it is also possible to base clustering on data bases of documents labeled according to predefined categories that are either hand-assigned (e.g. books in libraries) or automatically assigned by machine-learned document classifiers. Clustering of unlabeled data and assigning labels to found clusters [Cheeseman, 1996] can be contrasted with clustering of pre-labeled data to discover weaknesses in the predefined labeling. Chosen labels may be ambiguous; two labels may cover overlapping types of documents, and clustering can reveal problems of these types.

Association discovery

Other types of knowledge may also be mined from text through the analysis of argument structure. Basic elements of argument structure are the predicate relations between concepts such as 'leads to', 'is associated with', 'never occurs with', 'is a part of'. Given that concepts are discovered in the preprocessing and analysis phases, their relations can to some extent be computed. An example of knowledge mining using these relations is given by [Swanson, 1997], see Section 2.2.1, Text mining for science.

With a related technique, associative relations ('leads to', 'is correlated with') among identified concepts can be discovered. [Loh, 2000] gives some examples from a newspaper clippings collection with 358 texts:

drug traffic → politicians (confidence = 93,3%, support = 14 documents)

allowing the conclusion that the drug traffic problem reached the political sphere. Another rule:

loans AND education → politicians (confidence = 83,3%, support = 5 documents)

could lead to the conclusion that politicians are involved when loans and education are cited together.

The automated analysis of argumentative relations through both techniques may aid in the discovery of knowledge that is currently not present in one single document or known to one author or expert.

Knowledge representation

To make discovered information in text accessible, some form of representation is needed. The border between analysis and knowledge representation is not clear-cut, since displaying the raw results of for instance an association rule discovery process is a form of knowledge representation. However in most of the cases an extra step can provide much better insight and understanding.

Visualisation is particularly useful to provide insight in relationships between concepts or documents. Visualization might range from showing the terms with their frequencies to 3 D landscapes of concepts or documents.
Often it is very hard to visualize extended textual data in the more advanced 2 or 3D visualizations (such as SOMS: see http://websom.hut.fi/websom/, or browsing visualization tools [Wiesman, 1998], and it is a topic in many research projects.

Non-visual representations of knowledge in documents can take the form of ontologies or wordnets, in which the essential (informationally most salient) concepts in a collection of documents are represented by nodes, and arcs between nodes represent relations. Usually, but not necessarily, such networks of connected nodes have a hierarchical structure, in which the higher non-terminal nodes represent more general concepts. For example, consider the set of all web pages of all computer science departments at universities worldwide. High-level salient concepts represented as relatively high nodes in the discovered structure would be ‘courses’, ‘departments’, ‘research projects’, ‘students’, ‘professors’, etc., while at the lowest, most specific level concepts (nodes) would represent individual persons and particular courses.

This computer science example has been worked out fully in the Web-KB project [Craven, 1998], where a working knowledge base of US computer science was inferred on the basis of the raw collection of all web pages from computer science departments. Before this knowledge base, no such overview of the whole of US computer science was available; information is usually only accessible within an individual node. All pages in the raw collection were classified as being that of a staff member, student, etc. or representing a course, or a research project, or a department. Hyperlink structure on these pages was sub-

sequently used to establish relations among the nodes, where the found links represent relations such as 'is a member of', 'is student of', and 'is teacher of'. On a staff member's page, he or she would typically place links to the courses he or she teaches; a department would typically publish a linked list of its staff members, etc.

An important conclusion of the Web-KB project is that web documents offer easier handles for inferring ontological relational knowledge than non-hypertext documents, because of the hyperlink structure. Hyperlinks represent strong relational information, endorsed by the author who chooses to use them. With the continuous growth of the Internet, the current expectation is that text mining from web documents will provide the fastest discovery of new knowledge.

References

- Cheeseman, P., J. Stutz (1996). Bayesian Classification (AutoClass): Theory and Results. In: U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy. (eds.). Advances in Knowledge Discovery and Data Mining. AAAI Press/MIT Press
- Craven, M., D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, S. Slattery. (1998). Learning to Extract Symbolic Knowledge from the World Wide Web. Proceedings of the 15th National Conference on Artificial Intelligence (AAAI-98)
- Cucerzan, S., D. Yarowsky. (1999). Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence. Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora. pp90-99
- Hearst, M.A. (1999). Untangling Text Data Mining. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, ACL'99. pp3-10. ACL, New Brunswick, USA
- Joachims, T. (2002). Learning to Classify Text Using Support Vector Machines. Kluwer
- Kohonen, T. (1989). Self-Organization and Associative Memory. 3rd Edition. Springer Verlag, Berlin
- Loh, S, L.K. Wives. J.P. de Olivera. (2000). Concept-Based Knowledge Discovery in Texts Extracted from the Web. SIGKDD Explorations **2** (1)
- Mather, L.A., J. Note. (2000). Discovering Encyclopedic Structure and Topics in Text. KDD 2000 Workshop on Text Mining, Boston, MA
- Mladenic, D. (2001). Using Text Learning to Help Web Browsing. Proceedings of the 9th International Conference on Human-Computer Interaction (HCI International 2001), New Orleans, USA
- Rauber, A., D. Merkl. (1999). Mining Text Archives: Creating Readable Maps to Structure an Describe Document Collections, in Principles of Data Mining

and Knowledge Discovery. Proceedings PKDD'99, Springer Verlag, Berlin
- Salton, G. (1989). Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, Reading, USA
- Shankar, S., G. Karypis. (2000). A Feature Weight Adjustment Algorithm for Document Categorization. KDD 2000 Workshop on Text Mining, Boston, MA
- Shannon, C. (1948). A Mathematical Theory of Communication. Bell System Technical Journal **27**:379-423. pp623-656
- Steinbach, M., G. Karypis, V. Kuma. (2000). A Comparison of Document Clustering Techniques, University of Minnesota, USA. Technical Report #00-034, http://www.cs.umn.edu/tech_reports/
- Swanson, D., N.R. Smalheiser. (1997). An Interactive System for Finding Complementary Literatures: a Stimulus to Scientific Discovery. Artificial Intelligence **91**:183-203
- Wiesman, F., A. Hasman. (1997). Graphical Information Retrieval by Browsing Meta-Information. Computer Methods and Programs in Biomedicine **53** (3):135-152
- Yang, Y. (1999). An Evaluation of Statistical Approaches to Text Categorization. Journal of Information Retrieval **1** (1/2):67-88

6

6.3 Visualization and Interaction

6.3.1 Explorative visualization

Jeroen Meij

Perception

Visualization is an effective way of representing large quantities of data. While it is extremely difficult to see structure in the table of a database, often a graphical representation of the same table can be very easy to interpret. [Ware, 2000; Vesanto, 2001] stress the importance of pre-attentive processing, which is performed on the visual input before the conscious processing takes place. Pre-attentive recognition does not increase with the number of distractors, and is much faster (<10 ms) than conscious processing (>40 ms). Therefore, presenting the data with pre-attentive features has a lot of advantages. Table 1 lists the most important pre-attentive features.

Table 1
Pre-attentive features [Ware, 2000; Vesanto, 2001].

Position	Color	Shape	Motion	Texture
2D position	Hue	Size	Flicker	Orientation
Stereoscopic depth	Intensity	Elongation	Direction of motion	Size
		Orientation		Contrast

Explorative visualization

In this chapter a short overview of important visualization methods is given, with the accent on explorative techniques. The aim is to discover new information from the data, not just presenting data nicely or confirming visually what is already known. Some goals of explorative visualization:

- To show many values in one viewable area.
- To allow objects to be compared to each other.
- To allow the recognition of groups (clusters).
- To reveal relations between data dimensions.
- To allow a model or overall pattern to be compared to the data.

A process for explorative visualization is given in Figure 1.

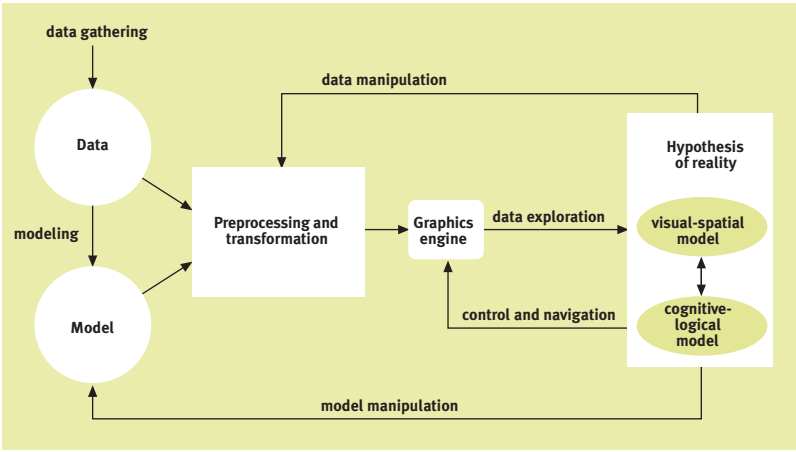


Figure 1
Visual data mining loop [Ware, 2000; Vesanto, 2001]. The model can be verified and refined in iterative loops.

Visualization itself can be divided into rendering and navigation, which will be discussed below. The rendering part is often also called visualization. For clarity we will stick to rendering in this section.
This overview cannot be exhaustive, considering the hundreds of methods that have been described in literature. After this overview Section 6.3.2 discusses one explorative method in more detail, Self-Organizing Maps or SOM's [Kaski, 1997 (CD-rom)]. Section 6.3.3 discusses interaction in virtual reality environments.

Navigation

In any interactive visualization, some form of navigation is required. [Chalmers, 1999] distinguishes three types of navigation in visual representations:

- *Spatial*: This mainly relates to data that is mapped to represent the original 2D or 3D sources, as observed in physics and geography.
- *Semantic*: This refers to the metaphorical use of space, with assumptions like close = similar, and high = important.

– *Social*: Social navigation uses the past activities of others to facilitate navigation, examples are used of stored 'wear' patterns from reading or editing by others. Collaborative filtering and collaborative paths may also be used.

Navigation is very important for the understanding of the data. These techniques allow the user to select different viewpoints, projections, filtering methods etc. and see the changes in the visualization.
An advanced form of this interaction is discussed in 6.3.3, the section on Interactive tools and Virtual reality.

Rendering

Five main categories of data exploration rendering can be used to group the techniques (largely based on [Ankerst 2001; Keim, 1997]:

1 Geometric.
2 Icon-based.
3 Pixel oriented.
4 Hierarchical.
5 Graph-based.

These rendering techniques can be complemented with distortion techniques and interaction techniques (navigation). In general distortion will focus the user's attention to a specific part of the rendered data, using perspective projections, lens-like effects and transformations.

Geometric techniques

For n dimensional data, we can build a matrix of scatter plots for each combination of data pairs ($n^2/2$-n plots). From these plots we can see possible relations between the variables.
Contour (height) maps also visualize the relationship between two variables, but are based on lines rather than points. These lines show constant values, combining categories with quantitative data, for instance by using color for different categories [Minty, 1996].
Related to geometric techniques are volumetric techniques. From the height map mentioned above, it is relatively easy to create a three dimensional landscape that can be drawn from all angles. Many techniques are available to make this landscape easier to interpret: adding shadows, reflections, textures (for instance for different soil types, or dense areas). Although these techniques are mainly used to visualize volumetric data (geographic, product geometry, weather), many other uses exist. Some examples are web search results and document analysis data.
Parallel coordinates is a technique in which every variable is given its own axis all axes are at equal distances from each other, and scaled to each variable's

minimum and maximum value. A line is drawn for each record through all axes, resulting in a graph with many converging and diverging lines.

Icon-based techniques

Certain characteristics of the data can be mapped to meaningful geometric shapes (icons). In this way, the icons can be placed in simple graphs where the X, Y, Z coordinate of the icon indicates two or three important variables, while the icons themselves are representing other variables. Well-known icon-based techniques are Chernoff faces, Stick figures and shape- or color coding. Chernoff faces are suitable to recognize trends in data, where up to twenty variables can be coded in a face. Plotting a series of faces can reveal trends in one or more variables (see Figure 6).

Pixel oriented techniques

Each attribute value is represented by a colored pixel, the value ranges of the attributes are mapped to a fixed color map. The attribute values for each attribute are presented in separate windows.

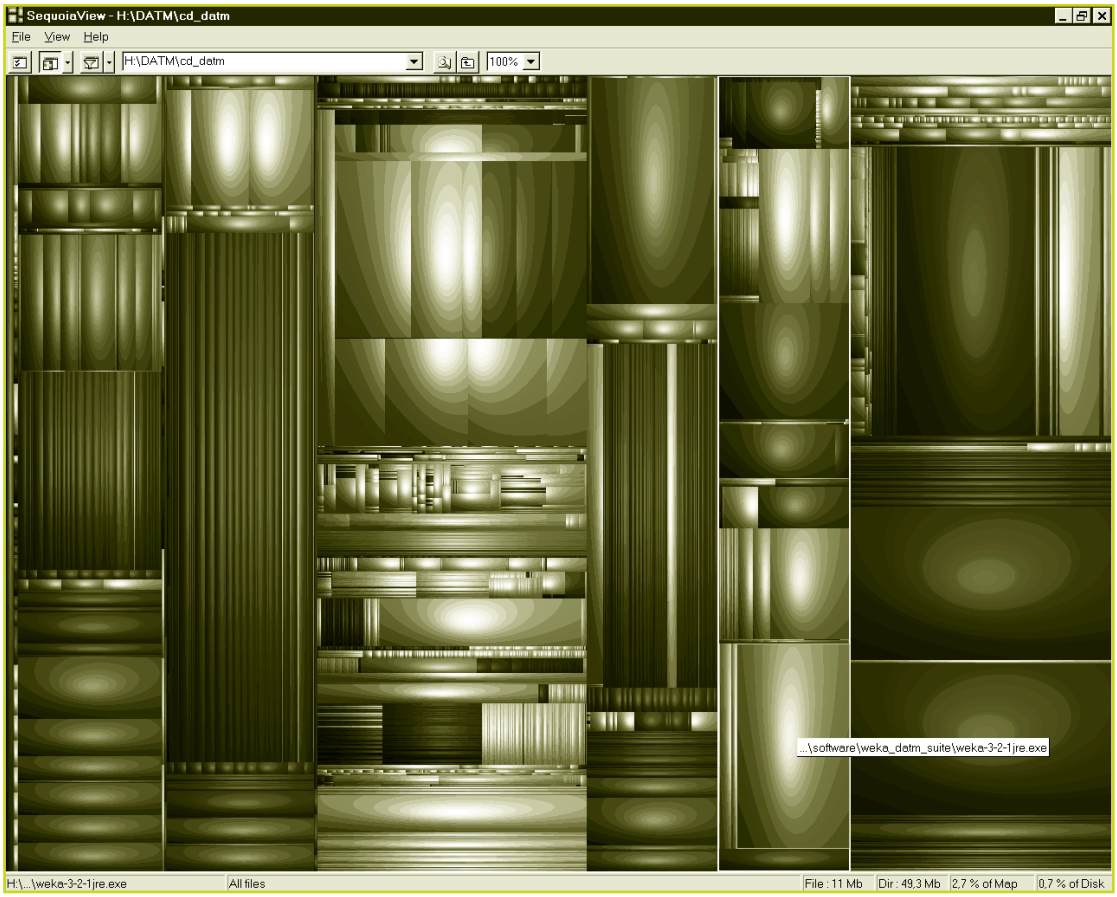


Figure 2
Treemap of a directory structure. The highlighted part is the software directory on the CD-rom.

Hierarchical techniques

The data is visualized using a hierarchical partitioning into subspaces. A well-known example is the Treemap [Shneidermann, 1998], which visualizes every subspace proportionally to its size, but Venn diagrams also belong to this category. Treemap software is included on the CD-rom (Sequoiaview) [Wijk, 1999].

Graph-based techniques

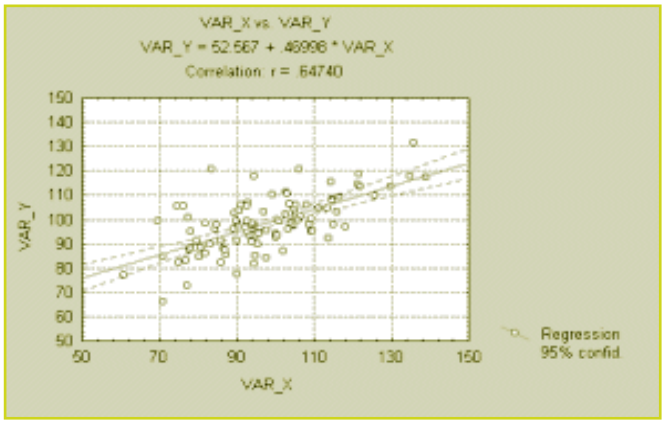
Typically a graph consists of lines and nodes, lines can be straight or curved, nodes can be with or without a surface, volume or label. Graphs can be drawn in 2D or 3D.

Data dimensionality

The number of dimensions in the data set is an important factor for the selection of a visualization method. Below we will describe some common visualization techniques, grouped according to the dimensionality they can handle.

Two or less dimensions

A widely used technique for 2 dimensional data is the x-y plot of two variables, the scatter plot, which visualizes the data points in X and Y direction. From the shape and direction of the group(s) of points, conclusions can be drawn about the relation between X and Y (linear, log, etc.).
Contour (height) maps also visualize the relationship between two variables, but are based on lines rather than points. These lines show constant values, combining categories with quantitative data, for instance by using color for different categories [Minty, 1996].



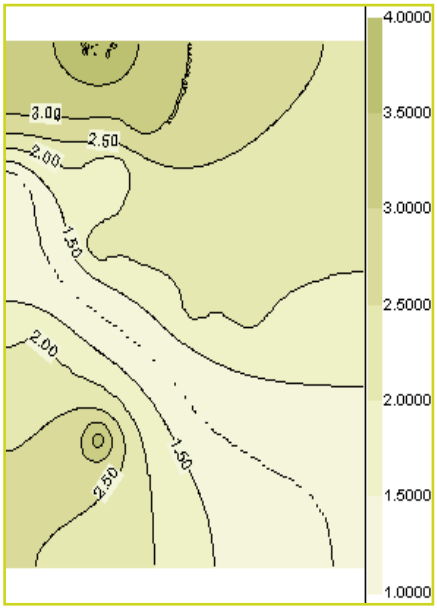


Figure 3
Scatter plot (left). Source: [Statsoft,2001] and contour map. Source: [3Dfield, 2001].

Three dimensions

Obviously, the 3-dimensional plot is widely used to visualize 3D data, but is only easy to interpret if the user is able interact with the projection (rotate, zoom, etc.), or the projection enables depth perception by stereoscopy or other techniques.

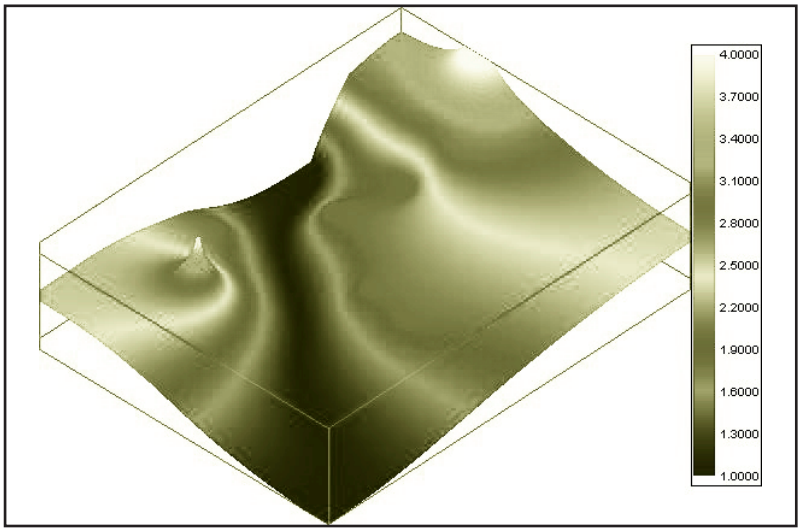
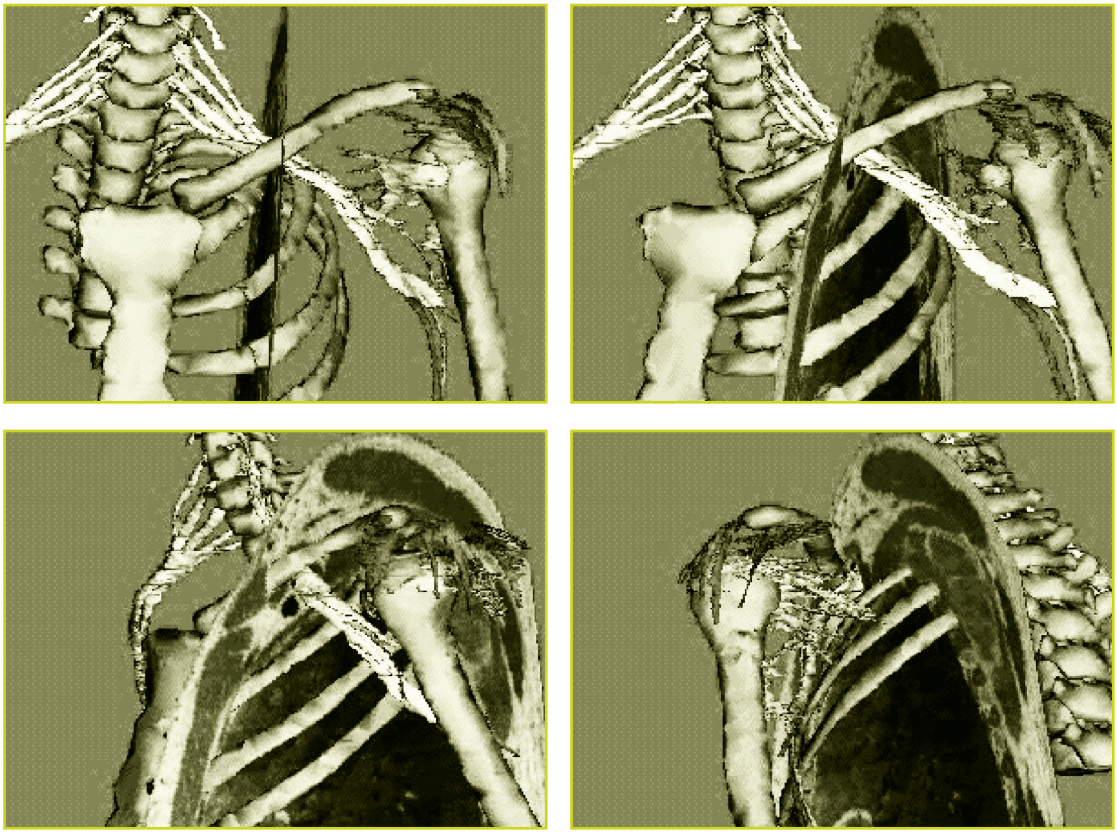


Figure 4
3 dimensional plot of the contour map of Figure 3. Source: [3Dfield, 2001].

Figure 5
Interactive combined 3D and 2D projection from the Visible Human dataset [NLM, 2002; Gold, 2002]. Source: [EPFL, 2002].



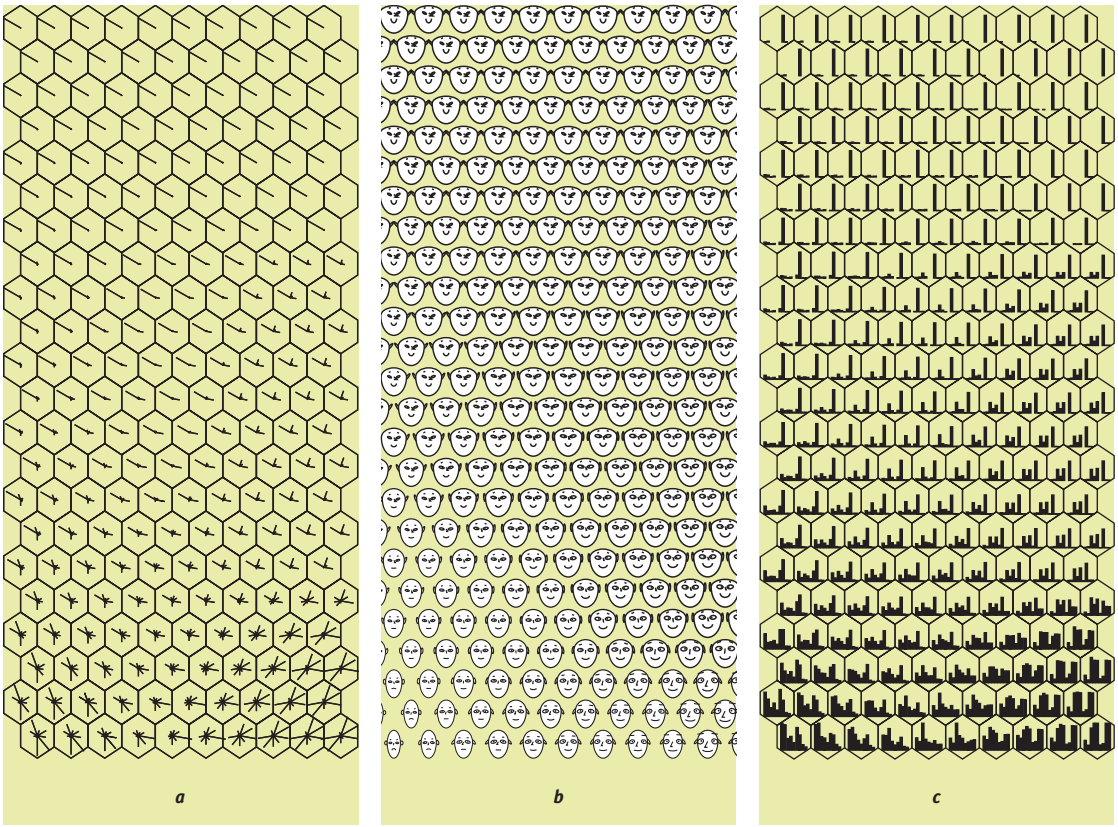


Figure 6
Visualizations of 9 dimensions (vectors) grouped in an organized (SOM) grid. Each unit is visualized as a glyph:
a *Fan plots,*
b *Chernoff faces,*
c *bar plots.*
[Vesanto, 2001].

Three to 8 dimensions

We can visualize up to 8 dimensions by multiple coding of dimensions into color, motion, shape or texture.

Ten or more dimensions

For 10 or more dimensions, we have to

- Add non pre-attentive features. This will extend the range to 29 dimensions, but dimensions above 9 will be less intuitively interpretable. Well-known methods are glyphs (Chernoff faces, Stick figures, whiskers) or using embedded dimensions.
- split the data into multiple visualizations linked by position, color, lines or motion [Buja, 1991; Vesanto 2001] (see Figure 7), or
- remove redundant or less important dimensions;
- combine multiple dimensions into one new dimension.

Figure 7 (opposite page)
Multiple linked visualizations, Small multiples of the same set of data. In the component planes figure (a), linking is done by position. Each subplot corresponds to one variable, and each dot in each subplot to one data sample. The coordinates are from the PCA-projection of the data.

Dimension reduction techniques

To reduce the dimensionality of data — to enable useful visualization — a variety of data reduction techniques can be used. Well-known techniques include [Ankerst, 2001]:

The last component plane shows the color coding of the map, and thus links the component planes with (b-d). In the time-series plot (b) the data for Tuesday 6:00 to 22:00 is depicted. The samples have been colored using the color coding. The scatter plot matrix (c) shows the pair wise scatter plots of all variable pairs. The objects in the scatter plots have been linked together using color. On the diagonal, the histograms of individual variables are shown. In the parallel coordinates visualization (d) each vertical axis corresponds to one variable. Each horizontal line corresponds to one object such that linking is done explicitly using lines. Like in (a), similar objects have been encoded using color. Figures (a), (c) and (d) all show all of the data, and can be used to detect correlations between variables. [Vesanto, 2001].

- Factor analysis [Harman, 1967].
- Subspace methods, including Principal Component Analysis (See Section 6.2.4).
- Multidimensional scaling (See Section 6.2.5).
- Fastmap [Faloutsos, 1995].

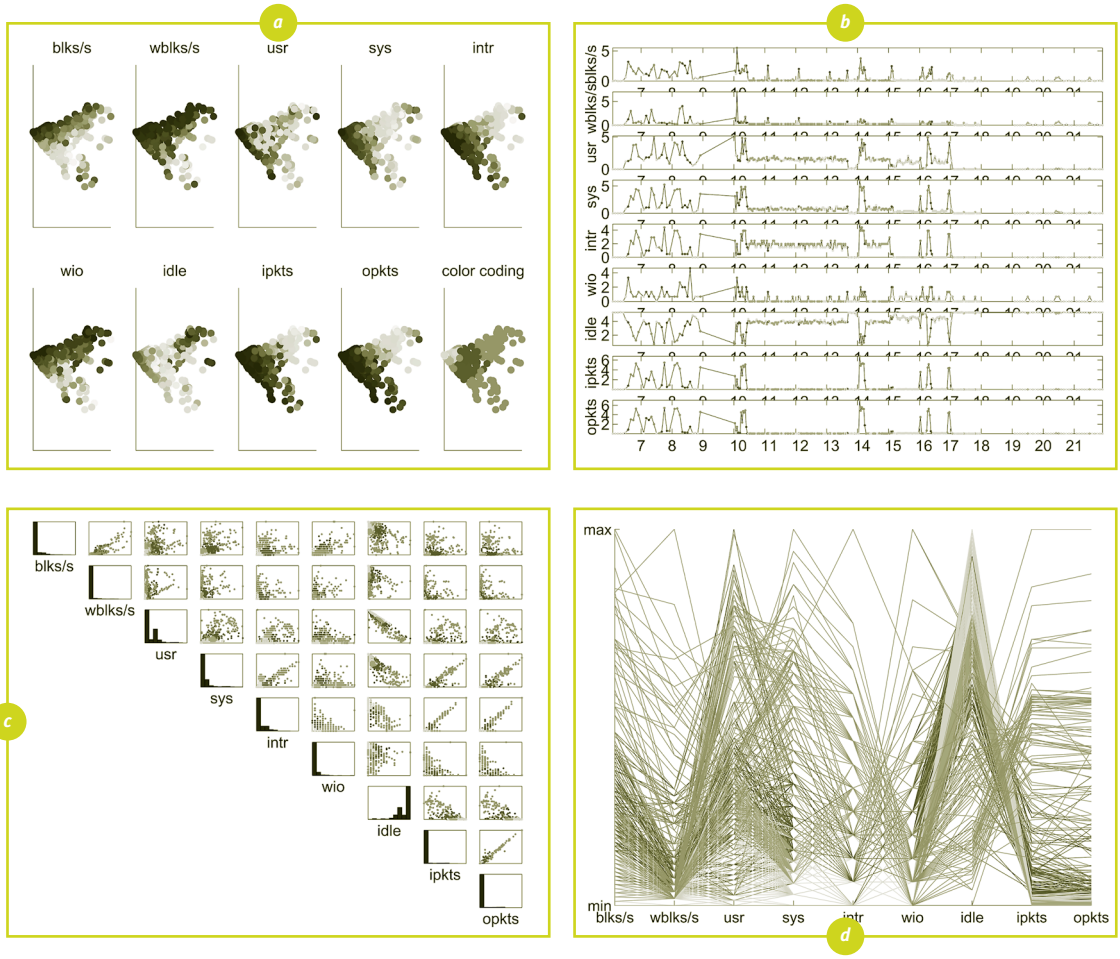
See also Section 2.3.11 for examples of dimension reduction in behavioral science.

Visualizing Data Mining Results

Relatively new is the development of special techniques for the visualization of the results of data mining.

For clustering, classification, association rules and text mining results, special visualization techniques are available. For a comprehensive overview see [Ankerst, 2001].

Some highlights are H-BLOB [Sprenger, 2000] for visualizing clustering results, and interactive mosaic plots [Hoffmann, 2000] for association rules, the latter being discussed in Section 6.2.12.



Acknowledgement

I would like to thank Juuha Vesanto for his advice and for allowing me to use his figures.

References

- 3Dfield. (2002). Contouring and 3D Surface Plotting Program. http://field.hypermart.net
- Ankerst, M., D.A. Keim. (2001). Visual Data Mining and Exploration of Large Databases. Tutorial Slides. ECML–PKDD 2001. http://www.afia.polytechnique.fr/CAFE/ECML01/visual_dm.html. http://www.afia.polytechnique.fr/CAFE/ECML01/T08.pdf
- Chalmers, M. (1999). Tutorial on Information Visualization, VLDB'99. http://www.dcs.gla.ac.uk/~matthew/
- EPFL. (2002). Visible Human Server. http://visiblehuman.epfl.ch
- Faloutsos C., K. Lin. (1995). Fastmap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets. Proceedings ACM SIGMOD International Conference on Management of Data, San Jose, CA. pp163-174
- Gold. (2002). Segmented and Classified Data from Gold Standard Multimedia. http://www.gsm.com
- Harman, H.H. (1967). Modern Factor Analysis. University of Chicago Press
- Hofmann, H., A. Siebes, A. Wilhelm. (2000). Visualizing Association Rules with Interactive Mosaic Plots. Proceedings ACM SIGKDD International Conference On Knowledge Discovery & Data Mining (KDD 2000), Boston, MA
- Kaski, S. (1997). Data Exploration Using Self-Organizing Maps. PhD Thesis. Helsinki University of Technology. Acta Polytechnica Scandinavica: Mathematics, Computing and Management in Engineering **82**
- Keim, D.A. (1997). Visual Techniques for Exploring Databases. KDD 1997
- NLM. (2002). National Library of Medicine Visible Human Project. Professor Ackerman. http://www.nlm.nih.gov/research/visible/
- Shneiderman, B. (1998). Treemaps for Space-Constrained Visualization of Hierarchies. Human-Computer Interaction Lab (HCIL) at the University of Maryland. http://www.cs.umd.edu/hcil/treemaps/
- Sprenger, T.C., R. Brunella, M.H. Gross. (2000). Hierarchical Visual Clustering Method Using Implicit Surfaces. ETH Zurich, Tech Report. http://graphics.ethz.ch
- Vesanto, J. (2001). Visualisation Methods. IDA Spring School Tutorial, Palermo. http://www.cis.hut.fi
- Wijk, J.J. van, H. van de Wetering. (1999). Cushion Treemaps. In: G. Wills, D. Keim. (eds.), Proceedings 1999 IEEE Symposium on Information Visualization (InfoVis'99). October 25-26. IEEE Computer Society. pp73-78

6.3.2 Self-Organizing Maps, a visual exploration tool

Martijn Schuemie[1], *Jan van den Berg*[2], *Roger P.G.H. Tan*[3]

Introduction

We humans are good at detecting visual patterns. Driving a car, scanning the headlines of newspapers or noticing that a single tile is out of alignment in your bathroom; during most of our daily activities we rely heavily upon the information we can extract from what our eyes can see. Our affinity for visual imagery can be used to effectively and efficiently convey information. This has been done for ages. Graphs and charts have been used to give better insight into data than is possible with numbers alone.

This paper is about Self-Organizing Maps (SOMs), an algorithm that can be used to turn data into something we can see and understand: a two-dimensional map. Tools like SOMs are invaluable in data mining. They provide the means to make high-dimensional data understandable and, perhaps more importantly, explorable for the user. These visual exploration tools give the opportunity to first get a look and feel of the complex data; they do not require the user to ask the right questions right away.

First of all, this paper will elaborate on what SOMs are and how they work, followed by some examples of SOMs applied in datamining and closing with a look at the future of visual exploration tools.

Self-Organizing Maps

The Self-Organizing Map algorithm was introduced around 1982 [Kohonen, 1982]. Since then, several thousands of scientific papers have been written both on applications of SOMs and on mathematical analyses of the algorithm (for an overview, we refer to [Kaski, 1998]. In addition, several textbooks were published containing the theory and or applications developed using a SOM. Last but not least, various software packages have been developed to facilitate the application of the SOM algorithm (for an overview of these, see [Deboeck, 1998], Chapter 13).

Basically, a Self-Organizing Map consists of a grid of units. This usually two-dimensional grid is put in a high-dimensional data space, also called the input space. Our goal is to infer 'knowledge' from the input space, i.e. we want to understand the structure of this space. By putting the grid in the input space, every two-dimensional unit of the grid corresponds — at the same time — to a high-dimensional 'reference vector' in the given data space.

To illustrate, the height, weight and age of a person can be used respectively as *x*, *y*, and *z* coordinate of a point in a three-dimensional input data space. Figure 1a shows a number of samples of another population plotted in some three-dimensional input space. In an attempt to map this data onto a two-dimensional

1 Drs M. Schuemie,
m.j.schuemie@its.tudelft.nl,
Delft University of Technology,
Faculty of Information Technology and Systems, Delft,
The Netherlands,
http://is.twi.tudelft.nl/~schuemie/

2 Dr Ir J. van den Berg,
jvandenberg@few.eur.nl,
Erasmus University Rotterdam,
Faculty of Economics, Department of Computer Science, Rotterdam,
The Netherlands,
http://www.few.eur.nl/few/people/jvandenberg/

3 Drs R.P.G.H. Tan,
tan@mediaport.org,
Robeco Group N.V., Quantitative Research Department, Rotterdam,
The Netherlands

Figure 1a
Points in a data space.

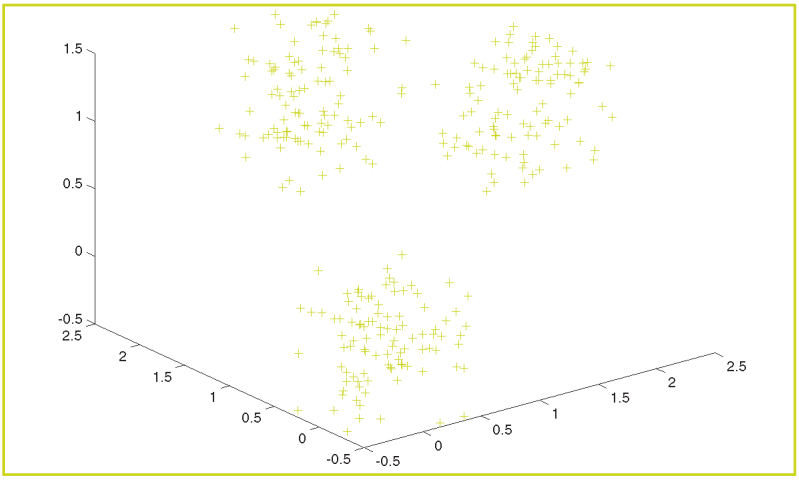


Figure 1b
Linear initializations of the reference vectors of a SOM.

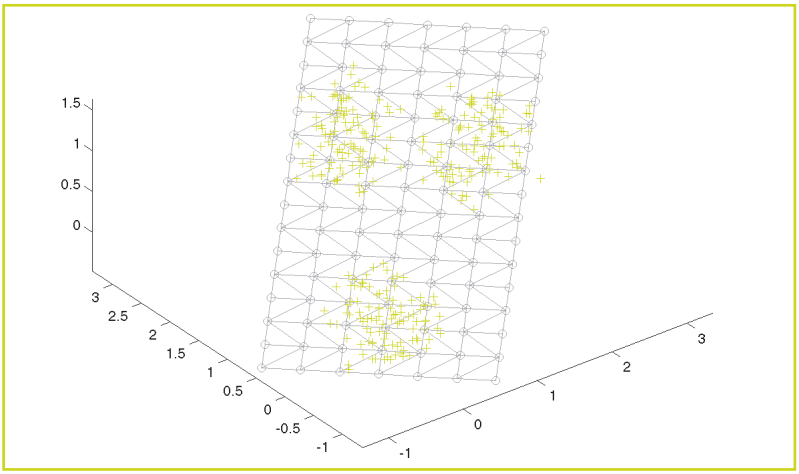
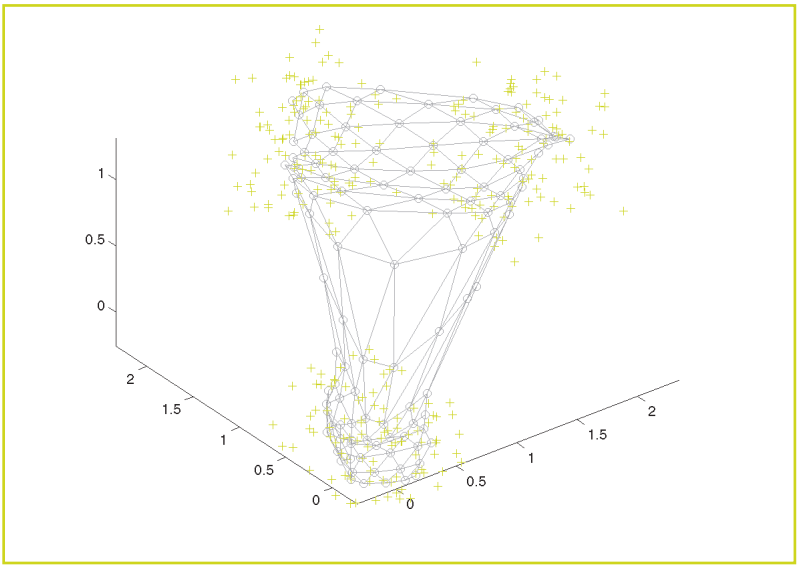


Figure 1c
Reference vectors after applying the SOM algorithm.



map, the reference points of the SOM-units are placed in this space. In the beginning these points can be placed at random or, to speed up the algorithm, using some form of linear initialization as depicted in Figure 1b. SOMs use an iterative algorithm to move the reference points from the initial state to a state where each reference point is close to a cluster of data points in the way that has been shown in Figure 1c.

To a certain extent, the structure of the three-dimensional space is mapped on the two-dimensional SOM: the distribution of the reference points of the SOM mimics that of the original data. In addition, these reference points are also part of the two-dimensional grid and linked together as shown in Figure 1b and 1c. In this way, any data point in the input space, even a new one, can now be represented on a unit in the grid, simply by finding the reference point closest to that data point. The result is a two-dimensional map of the three-dimensional space. The most important feature of this map is that it is topology-preserving: points that are near each other in the input space are also near each other on the map. If we look at Figure 1c, we see that there are few units with reference vectors in between the clusters. The result on the map will be that units on the edge of one cluster are only a few units away from units on the edge of another cluster, even though in the data space the distance between them is relatively big. In order to visualize the distances between the reference vectors in the original input space, a color-coding is introduced in the map as shown in Figure 2. Here the darker colors indicate a large distance between reference vectors and bright colors indicate a short distance.

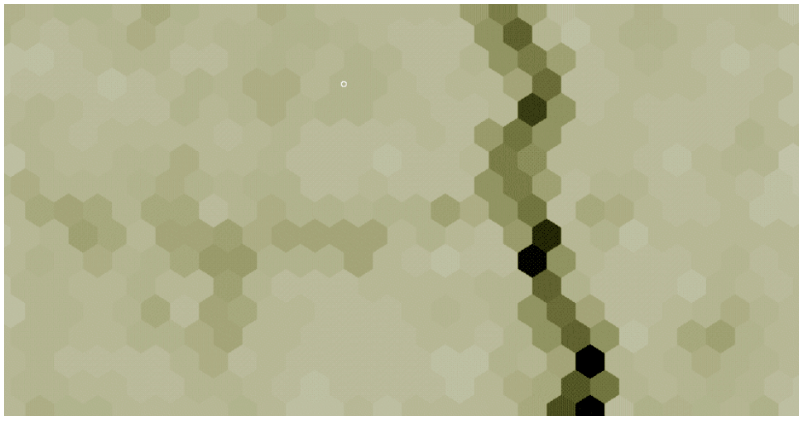


Figure 2
Two-dimensional map of the three-dimensional space with color-coding.

On the map we can now clearly see the distinction between the lower cluster in the data space (on the right of the map) and the two top clusters (on the left). The distinction between these last two is somewhat vaguer, since these clusters are not too far apart in the input data space.
Additionally, other color-coding schemes can be applied. For instance, a coding

based on the value of one of the components in the input space can provide the user insight in the distribution of this one characteristic over the data space. A good example of the use of such 'component planes' can be found in Section 3.2.2 of this book.
The example used here is a relatively simple one; a three-dimensional space is still comprehensible and the clusters already lie almost in the same linear plane, making it also possible to solve the problem using a simple linear mapping.

Figure 3a and 3b show a — in a certain sense — more complicated example where linear methods such as linear regression fall short. The input-space is two-dimensional here, while the grid (or map) is simply one-dimensional this time. The three clusters do not lie on the same line, and the regression model

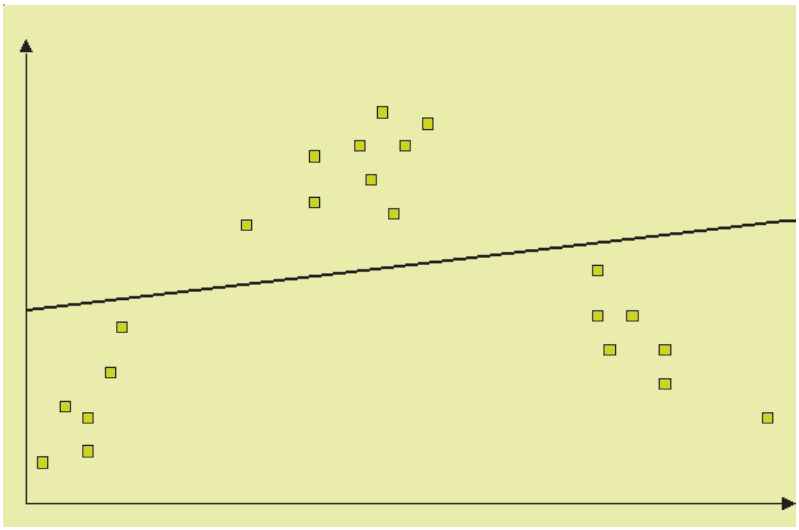


Figure 3a
Linear regression with a non-linear data set.

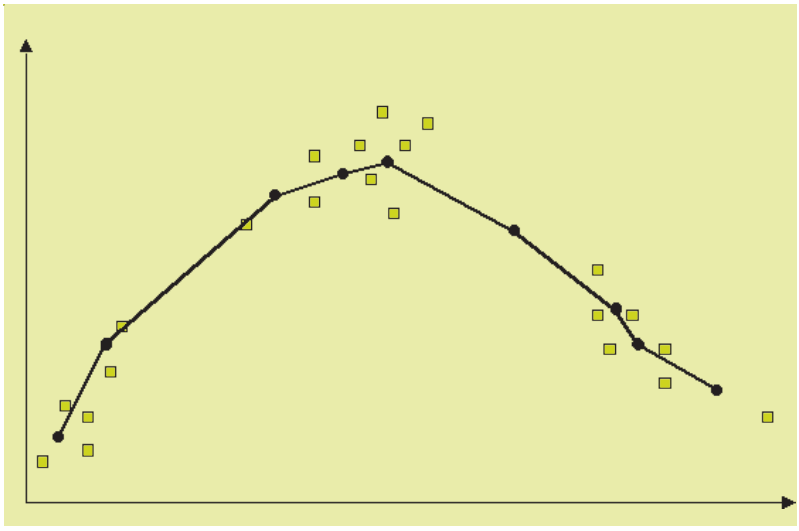


Figure 3b
A one-dimensional SOM with the same non-linear data set.

depicted in Figure 3a retains very little of the information in the original data. In Figure 3b, we see a one-dimensional SOM plotted in the two-dimensional space, providing a better fit for the data. In the SOM, it is still clear that medium *x* values are related to higher *y* values and that both lower and higher values of *x* correspond to lower values of *y*.

It can be seen that SOMs can make a non-linear projection of a higher dimensional space onto a space with lower dimensionality, while preserving topology. Sometimes this output space is used as input for other algorithms, for instance to classify input data. Most often, however, SOMs are used as an interface to the user. The often two-dimensional map, when provided with appropriate labels, can provide the user with unique insights into the data. The next paragraph will show some applications where the SOM is used specifically to communicate complex information to the user.

Examples of SOM datamining applications

Information Retrieval using ACS-WEBSOM

The field of Information Retrieval (IR) is concerned with enabling users to find specific information in large collections of documents. Good examples of IR-systems are search engines on the Web such as Altavista and Yahoo, and catalogue systems in libraries. Most such systems require the user to formulate a query, and words in this query are matched against words in the document title, author name, abstract and or keyword list. This requires the user to have a fairly good idea of what to look for already. The user must already be somewhat familiar with the domain and the terminology used in it.

A more visually oriented alternative is offered by algorithms such as the WEBSOM algorithm [Honkela, 1998] and our own extension, the ACS-WEBSOM algorithm [Berg, 1999] The notion underlying these systems is that documents can be characterized by the concepts addressed in them. Concepts are represented by clusters of words. The clusters themselves are found using a SOM and an additional learning and forgetting algorithm. For a detailed description of this process, we refer to [Schuemie, 1998].

Having done this, documents can be represented by points in another high-dimensional space where each dimension represents a concept as found in the previous step. For example: if a document contains many words found in the word cluster of concept *x*, then it's *x*-coordinate will be relatively high. This high-dimensional document space contains all the documents, arranged in such a way that documents that address similar topics are found close to each other. Again a SOM is used to make this space, which usually has several hundred dimensions, available to the user.

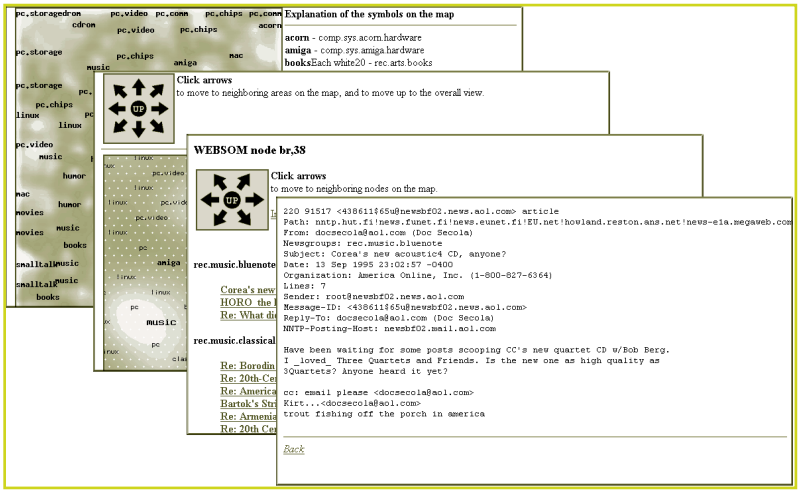


Figure 4
User interface of an IR-system using the WEBSOM algorithm. From left to right: 1. Starting screen with an overview of the collection. 2. Close-up of a specific region of the map. 3. A list of the documents related to a specific unit on the map. 4. Text of a document.

The resulting document map is labeled either manually or using a simple algorithm. If the map is very large, it can be made zoomable, as shown in Figure 4. The user can now explore the document collection by investigating the map. Clicking with the mouse on a unit on the map results in a list of the documents related to that unit. Documents in the same unit or nearby units are semantically related, allowing the user to find related information quickly. (An online demonstration of the WEBSOM is available at http://www.websom.hut.fi).

Other applications of the SOM

In Section 3.2.2 of this book another application of the use of SOMs is presented, namely in the area of assessing the creditworthiness of a company based on a large set of financial data concerning that company. In [Deboeck,1998], several other applications, mainly in the area of finance, are presented, ranging from a SOM clustering of more than 100 Scotch whiskies based on 72 different whisky features, to a SOM of about 50 mutual funds based on 15 fund features and intended to create a better basis for portfolio selection, and to a SOM intended to get a better understanding of the trends and patterns among today's emerging markets.

In [Kaski, 1998], an extended overview of SOM applications of all kind can be found. The articles cited concern applications in fields like machine vision and analysis, optical character recognition, speech analysis, signal processing and telecommunications, process control, robotics, mathematical problems, neurobiology, and more. In addition, many papers are devoted to mathematical analyses of the SOM algorithm and its extensions including convergence proofs, its relation to other mathematical fields like Markov-processes, energy-function formalisms, and Bayesian learning approaches.

In general, one may conclude that all SOM applications exploit one or more of its three 'basic properties' [Haykin, 1994]:

1 The Self-Organizing Map (represented by the grid points in the output space, each of which corresponds to a reference vector in the input space) provides a good approximation of the input space.
2 A Self-Organizing Map is topology preserving, which means that nearby grid points on the map correspond to nearby patterns in the input space.
3 Regions in the input space having a high density of data points are mapped onto larger domains of the output map. In this sense, a Self-Organizing Map reflects the statistics of the input data distribution.

Future

Around 20 years ago, the SOM-algorithm was invented, but it took quite a long time to make this algorithm available in an 'easy-to-use' way. Fortunately, nowadays user-friendly SOM–type software packages are available like Viscovery SOMine [Eudaptics, 1999]. This type of package has a visual interface simplifying the work of the data miner: (S)he is now able to concentrate on what (s)he is interested in, namely, on the discovery of knowledge. Besides creating the map (after automatically having executed several preprocessing tasks), Viscovery helps the data miner by converting the trained SOM into visual information. Several advanced data analysis tools are available within this package like data cluster search, numerical information retrieval including cluster statistics, data dependency evaluation, and more.

However, interpreting all emerging views on the data remains almost completely a task for the data miner. As in all areas of data mining, model validation is of great importance here. One could imagine that future software packages will become more intelligent by, for example, warning the user automatically when some structure has become visible which — after a more thorough statistical or other type of analysis — appears to be accidental instead of structural. In addition, techniques like cross validation [Haykin, 1994] for assessment of the generalization capabilities of a SOM outcome, might be performed in an automatic way to further support the future data miner.

Other improvements of the SOM can be achieved by further facilitating the human-computer interaction. Generating a SOM requires interaction with the user. Parameters for the SOM algorithm have to be set, and for example the user has to determine the basis for the color-coding, and this basis is often changed, when trying to interpret the map. This interaction currently requires the user to have a great deal of knowledge of the algorithm. Perhaps in the future the computer can aid the user in a dialogue fashion in refining and adapting the map in an interactive way to fit the user's need.

Also, the SOMs displayed to the user currently are all two-dimensional maps, while our perceptual system is based primarily on our environment, which is

three-dimensional. A 3D SOM could — in some applications — not only fit our way of viewing better, but could also be able to show intricacies of the input data not visible on a 2D map. How best to display a 3D map is still unclear however.

Summarizing, we think that SOMs will become a standard tool for data miners probably both in 3D and in 2D, available within standard office packages. In addition, we foresee that other machine learning algorithms based on the idea of visualizing information will be invented, since visual information is an efficient and effective way for rapid and correct data interpretation.

References

- Berg, J. van de, M. Schuemie. (1999). Information Retrieval Systems Using an Associative Conceptual Space. In: M. Verleysen. (ed.). Proceedings of the 7th European Symposium on Artificial Neural Networks (ESANN'99). pp351-356. Bruges, Belgium
- Deboeck, G., T. Kohonen. ((1998). Visual Explorations in Finance with Self Organizing Maps. Springer Verlag, Berlin
- Eudaptics. (1999). Viscovery SOMine. http://www.eudaptics.com
- Haykin, S. (1994). Neural Networks, A Comprehensive Foundation. MacMillan
- Honkela, T., K. Lagus, S. Kaski. (1998). Self-Organizing Maps of Large Document Collections. In: G. Deboeck, T. Kohonen. Visual Explorations in Finance with Self Organizing Maps. Springer Verlag, Berlin. pp.168-178
- Kaski, S., J. Kangas, T. Kohonen. (1998). Bibliography of Self-Organizing Maps. (SOM) Papers: 1981-1997. http://www.cis.hut.fi/research/refs/
- Kohonen, T. (1982). Self-Organized Formation of Topologically Correct Feature Maps. Biological Cybernetics **43**:59-69
- Schuemie, M. (1998). Associatieve Conceptuele Ruimte, een vorm van kennisrepresentatie ten behoeve van informatie-zoeksystemen. Master Thesis. (Available in Dutch). Erasmus University Rotterdam. http://www.few.eur.nl/few/people/jvandenberg/masters.htm

6.3.3 Dynamic exploration environments

Robert Belleman[1], *Peter Sloot*[2]

Introduction: exploration environments

In many scientific computing problems, the complexity of both the simulation and the generated data is too vast to analyze analytically or numerically. For these situations, exploration environments provide essential methods to present and explore the data in a way that allows a researcher to comprehend the information it contains. Exploration environments combine presentation and interaction functions into one system to allow exploration of large data spaces. These data spaces may originate from data acquisition devices or represent results from computer simulations. In our research we discriminate between static and dynamic exploration environments.

In Static Exploration Environments (SEE), the data presented to the user is time invariant; once the data is loaded into the environment, the user is presented with a visual representation of this data. Interaction methods are provided to change the visualization parameters interactively in order to get the best view to gain understanding. The data itself, however, does not change (see Figure 1).

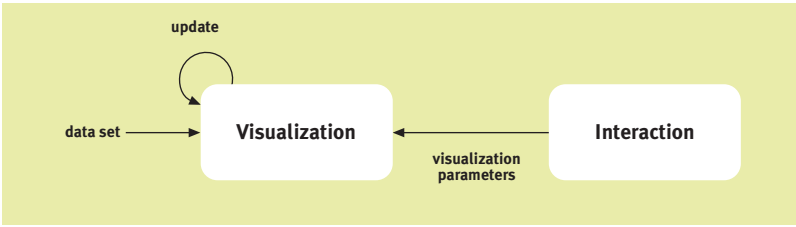


Figure 1
Schematic representation of a static exploration environment (SEE).

An important step towards a successful exploration environment is to involve the researcher in the presentation as much as possible, thereby increasing the researcher's level of awareness [Bryson, 1996a]. To achieve this, an exploration system needs the following, often conflicting capabilities:

- *High quality presentation*. The most common method to provide insight in large multidimensional data sets is to represent data as visual constructs that present quantitative and relational aspects to the observer in an comprehensible manner. Many scientific visualization environments are now available that provide means of efficiently achieving this [IBM, 1991; IrisExplorer, 1998; Schroeder, 1997; Upson, 1989].
- *High frame rate*. While the capabilities of modern graphical hardware allow increasingly complex images to be rendered with relative ease, the level of detail in the presentation should be minimized to avoid information clutter and achieve high frame rates (a compromise is often necessary). For an inter-

1 R.G. Belleman MSc,
robbel@wins.uva.nl,
The Universiteit van Amsterdam,
Faculty of Science, Section
Computational Science, Amsterdam,
The Netherlands

2 Prof Dr P.M.A. Sloot,
sloot@science.uva.nl,
The Universiteit van Amsterdam,
Faculty of Science, Section
Computational Science, Amsterdam,
The Netherlands

active exploration environment the visual frame rate should be at least 10 frames per second [Bryson, 1996b].

– *Intuitive interaction*. A prerequisite of a successful SEE is that a sufficiently rich set of interaction methods is provided that allows a user to extract both qualitative and quantitative knowledge from the data sets. An unfortunate side effect of increasingly richer sets of interactive methods is that user-friendliness is compromised, so careful consideration is required during user-interface design.
– *Real-time response*. Some delay will always occur between the moment a user interacts with a presentation and the moment that the results are visible. This is caused by low tracking rates of input devices, (re-computations, communication delays or temporary reduced availability of computational or network resources. To attain accurate control over the environment and to avoid confusing the user, the amount of lag in an exploration system should be minimized [Taylor, 1996].

Provided these capabilities are carefully considered, such environments are well suited for the exploration of static multidimensional data sets [Belleman, 1998]. Static exploration environments can be customized to observe iteratively updated data sets produced by 'living' simulations. When interaction with the simulation is also allowed, however, we speak of dynamic exploration environments and radically different considerations come into play, as we describe in the next section.

Dynamic exploration environments

Dynamic Exploration Environments (DEE) extend the previously described static model in such a way that the information provided to the user is regenerated periodically by an external process, in our case a computer simulation. Here, the environment is expected to provide (1) a reliable and consistent representation of the results of the simulation at that moment and (2) mechanisms enabling the user to change parameters of the external process (i.e. simulation) (see Figure 2).

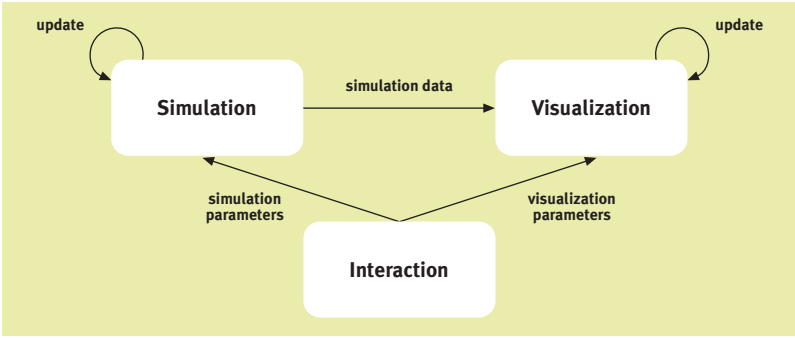


Figure 2
Schematic representation of a dynamic exploration environment (DEE).

Dynamic environments have additional requirements over static environments. For example, in static interactive systems, the interaction functions can be implemented inside the visualization environment, since the only interaction that takes place is with the visualization. In dynamic environments, interaction influences both the visualization and the simulation environment. Changing a static environment into a dynamic environment therefore requires that at least one module performs additional processing to service the interaction. Such a change makes these environment less suitable for use in other applications without significant modifications. In the sequel we address some of the functionalities required to develop generic dynamic exploration environments. We start with a brief description of the specific time management aspects in DEE and present a top-down description of the associated additional requirements in such systems.

Time management

An important issue in a DEE is time management. Time management deals with the exchange of time stamped information between components. For a DEE, the four most time demanding components are; the simulation environment, the visualization modules, the rendering layer and the explorer (i.e. the user).

Figure 3 and 4 show time frame diagrams illustrating the advancement of time, under two different time management strategies; lock-step and asynchronous. Time frames are illustrated by rounded boxes. The gaps between time frames on a same level represent the idle time of the component on that level. The gaps between time frames on neighboring levels represent the delays that occur between the time one component is done with a time frame and the next component starts working on it. These delays are delineated at the bottom of the Figure.

Figure 3
Time frames and delays in a lock-step interactive dynamic exploration environment.

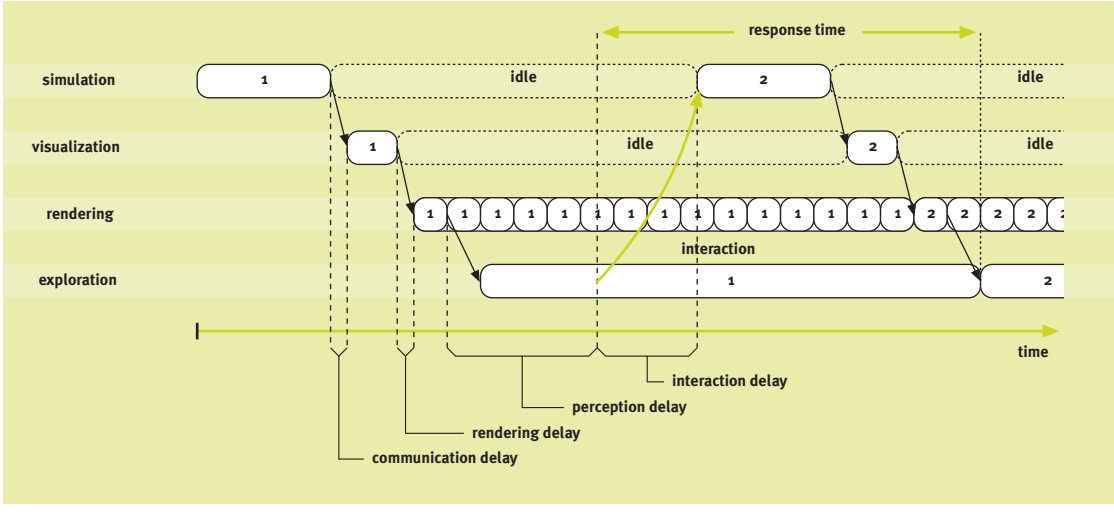


Figure 3 shows a time frame diagram in a lock-step interactive dynamic exploration environment. In this strategy, the simulation is allowed to advance, only if the user explicitly tells the environment it is alright to do so. While the user is exploring the results rendered by the graphics system, the simulation and visualization modules sit idle. In situations where a single simulation, visualization and rendering time frame takes a negligible amount of time, this strategy may be perfectly adequate, since the user will see the result of his interaction in short notice. However, if the these time frames are long, it may take a long time before the result of an interaction is shown. This often leads to an unusable environment, confusing the user.

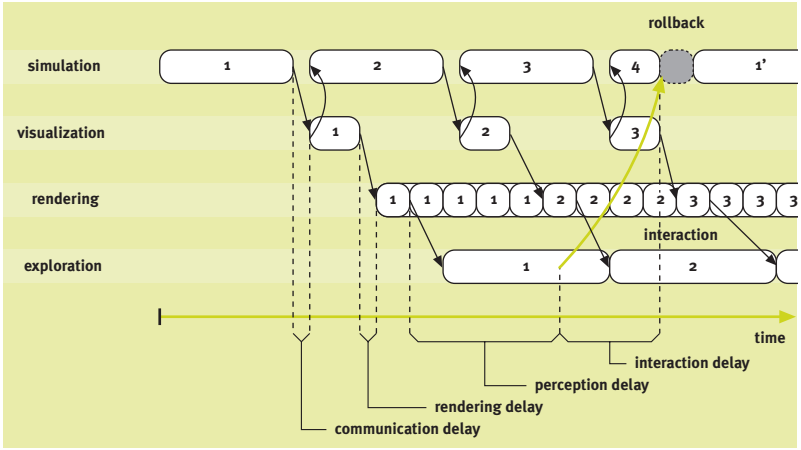


Figure 4
Time frames and delays in an asynchronous interactive dynamic exploration environment.

The time required before simulation updates are presented to the user (i.e. the length of a time frame on the exploration level) can be shortened by allowing the simulation to run asynchronously from the rest of the environment. Figure 4 shows a time frame diagram in an asynchronous environment. In an asynchronous system, different components are allowed to advance, when they have finished processing and communicating the current time frame. Different components may therefore execute at different time scales. In addition, these time scales are mostly non-deterministic because of hardware, software and human imposed delays. As a consequence, time delays occur as the output generated by one component cannot be accepted immediately by another component for processing. When components depend on the output of an increasing number of other components, the time frame that is processed by 'later' components fall further apart. This has a causality consequence for the user who explores the final component in the environment and therefore interacts with components at a much 'earlier' time compared to what is being processed by a simulation at the same wall-clock time. Time management is responsible for detecting and resolving this causality violation. Methods for resolving time causality problems have been investigated by [Overeinder, 1999].

Interaction

A DEE provides the opportunity to interact with a living simulation. This interaction can take any form; from typed input for simple types of interaction via graphical user interfaces to fully immersive virtual environments. The main feature of immersive environments over other graphical user interfaces is that user-centered stereoscopic images are presented to a user rather than visualization centered three-dimensional (3D) projections. User-centered stereoscopic images differ from projections on a flat screen in that slightly different pictures are generated for the left and right eye, dependent on the position of the viewer. This makes images 'pop out of the screen' and react to the user's movements[3], an important depth-cue to gain understanding in complex multidimensional structures.

A minimal requirement for interaction in an immersive Virtual Environment (VE, see also [Kaandorp, 1998]) is the availability of input devices that can be used to convey intention to the environment. The most common are sensor devices that measure the 6 Degrees Of Freedom (DOF) one has to move around in a 3D space (position and orientation). These sensors can be used to detect the proximity of a physical object (such as the user's hand) to virtual objects, so that the user can interact with them. Interaction with a virtual environment is a key issue, especially in an interactive simulation environment. The following types of interaction are deemed mandatory:

- *Object interaction*. An 'object' is defined here as a visual entity that is in the center of interest to an end-user. These objects are representations of data sets or simulation results, but can also be 'widgets' (menus, buttons, sliders, etc.). An object has attributes associated with it such as position, scale, level, state, etc. Object interaction is concerned with changing these attributes.
- *Navigation and way finding*. Navigation provides users with methods to move beyond the confinements of the VE's physical dimensions. Objects beyond the VE come into reach by moving the user towards them. Note how this concept places the user of the VE in the center of this type of interaction; the user is transported from one place to another, while the objects remain where they are. Way finding is a relatively new concept in VE applications and provides the user with a reference on where he is in a virtual environment [Elvins, 1997].
- *Probing*. Although visual presentations allow researchers to qualitatively analyze their data, an instrument for obtaining quantitative information from the visualization is a valuable asset. An architecture that allows researchers to probe visual presentations in order to obtain quantitative information is described in [Belleman, 1999].

3 This is commonly referred to as 'motion parallax'.

Coordination: intelligent agents

Since the various components in a DEE are independent processes, some means of coordination between them is required to allow a complex environment of this kind to be used. Especially in the case of interactive simulations, interaction involves not only the visualization element, but also the simulation part. For software engineering and efficiency reasons, it is reasonable to move the processing for those general interactions into independent modules, and let them be reusable to all components that need these interactions. One approach to this is through Intelligent Agents (IAs, see also Figure 5).

As are software modules with the capability of performing three functions: (1) perceiving state changes in the environment through the use of monitors, (2) taking actions that affect conditions in the environment and (3) reasoning to interpret perceptions, solve problems, draw inference, and determine actions [Hayes, 1995]. Agents execute autonomously, interfering minimally with the rest of the environment, apart from communicating with other agents or the user.

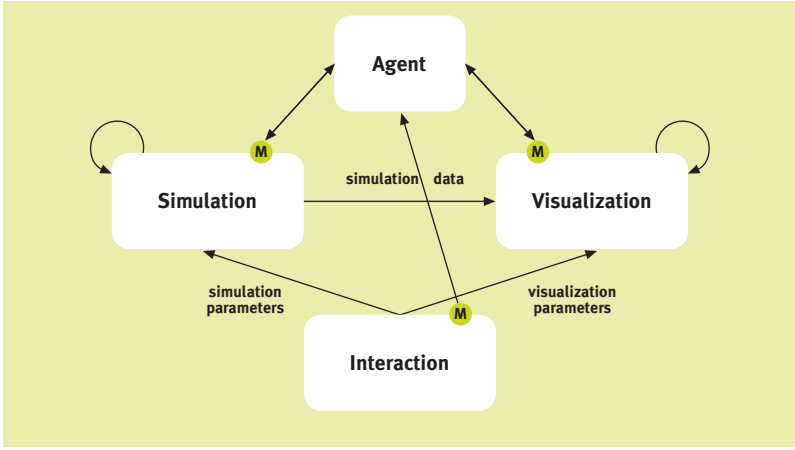


Figure 5
A DEE instrumented with monitors and an intelligent agent.

Feedback is generated when the agent has solved a problem, or has prepared suggestions based on the current running context. This feedback could be information to the user, or the information is sent to environment components. Depending on the permission settled beforehand, an agent could just provide feedback without affecting the working status of the environment, or make some changes in the environment based on its reasoning.
At present, agents are used to provide feedback to the user concerning the state of a simulation (e.g. accuracy of the simulation, time to completion, convergence rate) and user interaction (including speech recognition and synthesis). In the near future agents will be developed for feature extraction (e.g. determining the geometric skeleton of objects, detecting eddies in flow domains) and providing assistance (context sensitive help).

Distributed execution environment

Although the capabilities of modern computer systems are nearing the requirements for performing both simulation and visualization tasks on the same machine, some performance increase may be attained by running these tasks on dedicated computing platforms. For example, many simulation applications perform better on dedicated hardware such as vector processors, massively parallel platforms or other High Performance Computing (HPC) machinery, while state of the art graphical systems are now available that are well suited for the visualization tasks. Moreover, a decomposition of the environment into separate communicating tasks facilitates implementation and allows more control over the performance of the system as a whole (in Figure 5 each block can be considered a separate process or a combination of processes, possibly running on different systems).

Especially in the case of distributed environments, some means of job control is required that starts/stops the execution of the different components of the environment on the various computing platforms. In many organizations this system also needs to allocate the required resources prior to execution (for example in the case of batch execution systems). GLOBUS is one such software infrastructure for computations that integrates geographically distributed computational and information resources [Foster, 1997].

In distributed systems, components execute on different, possibly heterogeneous computing platforms. To be able to communicate data with each other, components provide access to their attributes, which can then be made available to other components. In heterogeneous computing environments the attributes often have to be converted into different representation formats. Furthermore, in many circumstances not all components in an environment will participate in a communication. For these situations a publication and subscription mechanism needs to be provided that limits communication to members of a restricted group.

Attribute ownership

The behavior of individual components in the environment is defined by one or more attributes (or parameters), which together define the state of that component. In a distributed system attribute changes (which can be considered to be events, for example as a result of user interaction) should only be performed by a component that owns the attribute to avoid race conditions. In some cases it may be necessary to transfer ownership so that attributes can be changed by other components (for example in a collaborative environment where multiple users manipulate the same components).

Runtime support system

From the considerations described in the previous sections, it becomes clear that a generic framework to support the different modalities is required. In our research we have chosen for the 'High Level Architecture' (HLA) as a suitable architecture for constructing a DEE.

HLA provides solutions to many of the problems and issues described in the previous sections. Specifically, HLA allows data distribution across heterogeneous computing platforms (including message groups), supports a flexible attribute publish/subscribe and ownership mechanism and offers several methods to do time management.

Vascular reconstruction: a case study

The design considerations described in the previous section cover the issues that are involved with building a DEE. The architecture is validated by analysis of a prototypical case study of simulated abdominal vascular reconstruction.

The application we have chosen as a test case combines visualization, simulation, interaction and real-time constraints in an exemplary fashion. By a detailed analysis of the spatial and temporal characteristics of the test case we attempt to recognize generic elements for the design of a computational steering architecture. We begin with a description of the test case.

Simulated abdominal vascular reconstruction

Vascular disorders in general fall into two categories: stenosis, a constriction or narrowing of the artery by the build-up over time of fat, cholesterol and other substances in the vascular wall, and aneurysms, a ballooning-out of the wall of an artery, vein or the heart due to weakening of the wall. Aneurysms are often caused or aggravated by high blood pressure. They are not always life-threatening, but serious consequences can result, if one bursts.

A vascular disorder can be detected by several imaging techniques such as X-ray angiography, MRI (Magnetic Resonance Imaging) or Computed Tomography (CT). Magnetic Resonance Angiography (MRA) has excited the interest of many physicians working in cardiovascular disease, because of its ability to non-invasively visualize vascular disease. Its potential to replace conventional X-ray angiography methods which use iodinated contrast has been recognized for many years, and this interest has been stimulated by the current emphasis on cost containment, outpatient evaluation, and minimally invasive diagnosis and therapy [Yucel, 1999].

A surgeon may decide on different treatments in different circumstances and on different occasions, but all these treatments aim to improve the blood flow of the affected area . Common options include thrombolysis where a blood clot dissolving drug is injected into, or adjacent to, the affected area using a catheter; balloon angioplasty and stent placement, which is used to widen a narrowed vessel by means of a inflatable balloon or supporting framework; or vascular surgery. A surgeon resorts to vascular surgery, when less invasive treatments are unavailable. In endarterectomy the surgeon opens the artery to remove plaque build-up in the affected areas. In vascular bypass operations, the diseased artery is shunted using a graft or a healthy vein harvested from the arm or leg.

The purpose of vascular reconstruction is to redirect and augment blood flow, or perhaps repair a weakened or a neurysmal vessel through a surgical procedure. The optimal procedure is often obvious, but this is not always the case, for example, in a patient with complicated or multi-level disease. Pre-operative surgical planning will allow evaluation of different procedures a priori under various physiologic states such as rest and exercise, thereby increasing the chances of a positive outcome for the patient [Taylor, 1998].

What is needed?

The test case described in the previous section contains all aspects of an interactive dynamic exploration environment that are of consequence in the construction of a generic dynamical computational steering architecture. Our aim is to provide a surgeon with an environment in which he or she can try out a number of different bypass operations and see the influence of these bypasses. The environment needs the following:

- An environment that shows the patient under investigation with his infliction. A patient's medical scan is 3D, so to obtain best understanding on the nature of the problem the surgeon should be able to look at his specific patient data in 3D, using unambiguous visualization methods.
- An environment that allows the surgeon to plan a surgical procedure. Again, this environment should allow interaction in a 3D world, with 6 DOF. The CAVE environment allows us to interact with 3D computer generated images using 6 DOF interaction devices [SARA, 1998; Cruz-Neira, 1993]. Note that visual realism is not the primary goal here; what is more important here is physical realism, and then only of particular issues in fluid flow, as discussed later[4].
- An environment that shows the surgeon the effect of his planned surgical procedure. As the aim of the procedure is to improve the blood flow to the affected area, the surgeon must have some means to compare the flow of blood before and after the planned procedure. This requires the following:

4 This in contrast to research projects towards virtual operating theatres that include the simulation of tissue deformation and realistic blood spills [Basdogan, 1999; Bockholt, 1999].

- a simulation environment that calculates pressure, velocity and shear stress of blood flowing through the artery;
- a visualization environment that presents the results of the simulation in a clear unambiguous manner;
- an exploration environment that allows the researcher to inspect and probe (qualitatively and quantitatively) the results of the simulation (e.g. means for performing measurements, annotate observations, releasing tracer particles in the blood stream, etc.).

All this should be interactive, or in other words, it should be fast enough in such that a surgeon does not have to wait for the simulation results.

Implementation of a dynamic exploration environment

Parts of the components mentioned in the previous section have already been implemented in the course of previous projects. Others require minor adaptations to fit into our dynamic exploration architecture. In the following subsections we will briefly discuss the current status of the visualization and exploration environment, the interaction environment, the simulation environment and the middleware that combines these together.

VRE: immersive static exploration

We have previously built a SEE called the Virtual Radiology Explorer (VRE [Durnford, 1999; Versweyveld, 1998]) which is capable of visualizing medical CT and or MRI data in 3D (see also Figure 6). 3D data sets acquired with CT or MRI are often displayed and evaluated from various perspectives or at different levels, including sets of single slices, stack mode (cine loop) interactive representation of sets of slices, or Multi-Planar Reformation (MPR) represented as single slices or interactive cine loops. Despite the increased possibilities of acquiring data, clinical use of 3D rendering has been hampered by insufficient computing capacity in the clinical environment.

An example of the clinical use of 3D rendering is simulated endoscopy[5]. Simulated endoscopy has several advantages over mechanical endoscopy (shorter acquisition times, increased patient comfort, higher cost-effectiveness, no complications of endoluminal instrumentation, field-of-view extending beyond the surface). In addition, simulated endoscopy can be used in virtual spaces that can not at all, or only after violation of normal anatomical structures, be reached by mechanical (endo)-scopy.

From a clinical perspective, there is a demand in community hospitals to make an environment available, suitable for the interactive rendering and interactive matching of, and switching between the above described data sets with an

5 Endoscopy is a diagnostic procedure where an instrument is used to visualize the interior of a hollow organ.

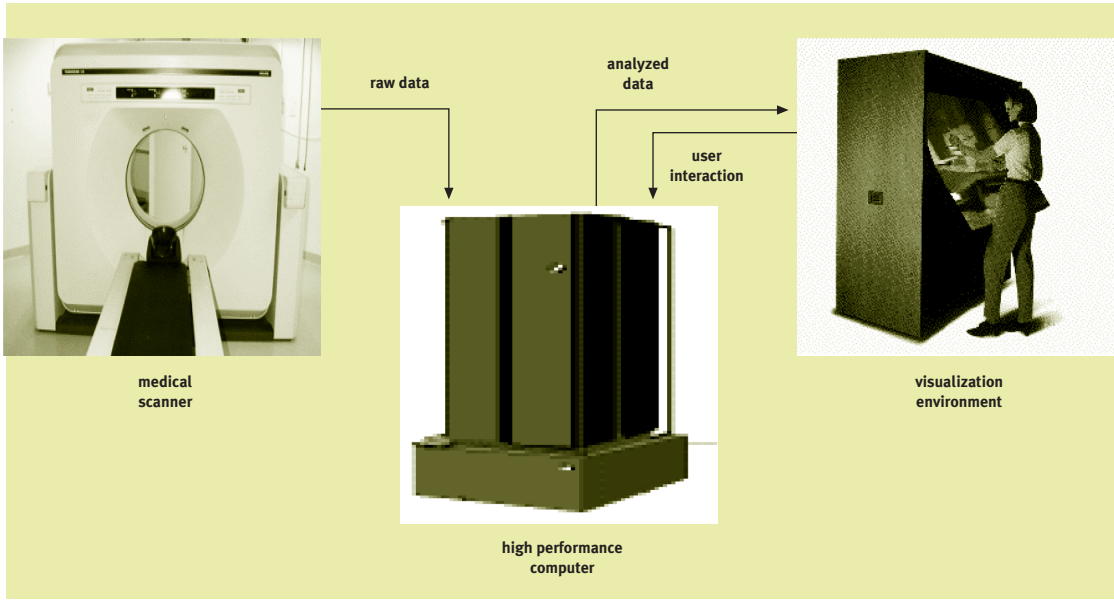


Figure 6
The VRE environment allows medical data from hospitals to be pre-processed on HPC systems for 3D visualization. High speed networking initiatives such as the GigaPort project [gigaport homepage, 2000] allow hospitals to make interactive use of HPC visualization techniques for patient diagnostics.

emphasis on simulated endoscopy. The VRE environment provides various such methods for the visualization of medical scans, including volume rendering using SGI's Volumizer [Volumizer homepage, 2000], surface rendering using VTK [Schroeder, 1997] and OpenGL [OpenGL homepage, 2000], interactive clipping and surface mapping techniques. Mechanisms have been added that allow the VRE environment to be run in a CAVE or on an ImmersaDesk. The ImmersaDesk allows the VRE environment to be used in the radiology department. Shown in Figure 7 is an isosurface representation of the abdominal aorta from an MRA scan. A geometric probing system (GEOPROVE, see [Belleman, 1999]) is used to perform measurements on this representation.

VRE+

VRE+ extends VRE with methods that allow dynamic exploration. Various methods are added to visualize the results of a simulation, while it is running. An intelligent agent system is integrated that constantly monitors a user's actions and provides feedback to the user. Currently, we have implemented a speech recognition agent, which enables users to control the environment using hands and voice simultaneously. A second agent monitors the position of the user, when using GEOPROVE and provides feedback on the accuracy of the measurements. Another agent monitors the state of the simulation environment and provides feedback on the state of the simulation.

For the planning part, the VRE+ environment is extended with means to 'draw' a surgical procedure using a 'grid editor', as described in the section on grid generation and editing.

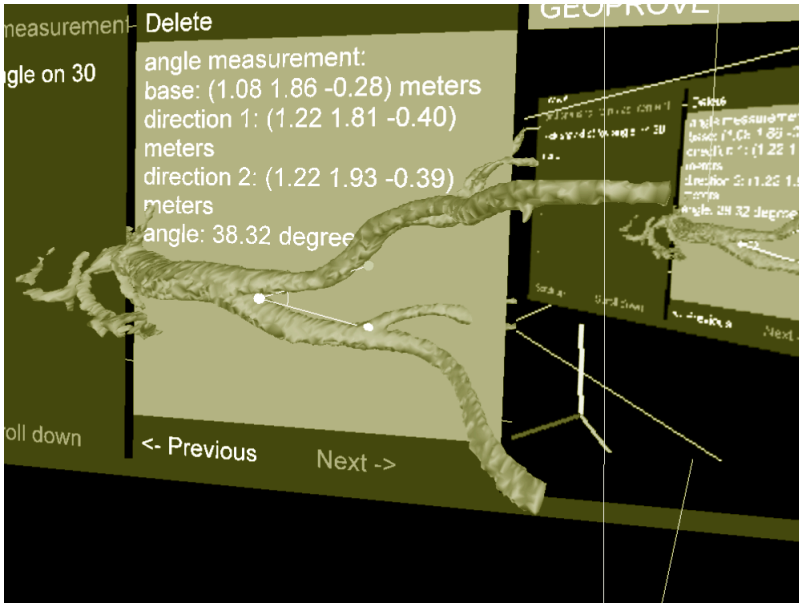


Figure 7
A snapshot of the VRE environment running in a CAVE. An isosurface representation of an abdominal aorta is shown obtained from an MRA scan. The panel shows the GEOPROVE environment, which allows measurements and annotations (such as the virtual 'snapshot' on the right) to be made from within the environment.

Fluid flow simulation: the lattice-Boltzmann method

The lattice-Boltzmann method (LBM) is a mesoscopic approach for simulating fluid flow based on the kinetic Boltzmann equation [Chen, 1998]. In this method fluid is modeled by particles moving on a regular lattice. At each time step, particles propagate to neighboring lattice points and re-distribute their velocities in a local collision phase. This inherent spatial and temporal locality of the update rules makes this method ideal for parallel computing [Kandhai, 1998]. During recent years, LBM has been successfully used for simulating many complex fluid-dynamical problems, such as suspension flows, multi-phase flows, and fluid flow in porous media [Koponen, 1998]. All these problems are quite difficult to simulate by conventional methods [Kandhai, 1999a; Kandhai, 1999b]. The data structures required by LBM (Cartesian grids) bear a great resemblance to the grids that come out of CT and MRI scanners. As a result, the amount of preprocessing can be kept to a minimum, which reduces the risk of introducing errors due to data structure conversions. In addition, LBM has the benefit over other fluid flow simulation methods that flow around (or through) irregular geometries (like a vascular structure) can be simulated relatively easily. Yet another advantage of LBM is the possibility to calculate the shear stress on the arteries directly from the densities of the particle distributions [Artoli, 2000]. This may be beneficial in cases where we want to interfere with the simulation while the velocity and the stress field are still developing, thus supporting fast data updating given a proposed change in simulation parameters from the interaction modules.

Lattice-Boltzmann grid generation and editing

As mentioned earlier, the basic structure of the grids used in LBM bear great resemblance to the medical scans obtained from a patient. To convert the medical scans into LBM grids, the raw data from the medical scanner is first segmented so that only the arterial structures of interest remain in the data set (see also Figure 8). The contrast fluid injected into the patient in a MRA scan provides sufficient contrast in the vascular structures to do this quite efficiently. The segmented data set is then converted into a grid that can be used in LBM; boundary nodes, inlet nodes and outlet nodes are added to the grid using a variety of image processing techniques.

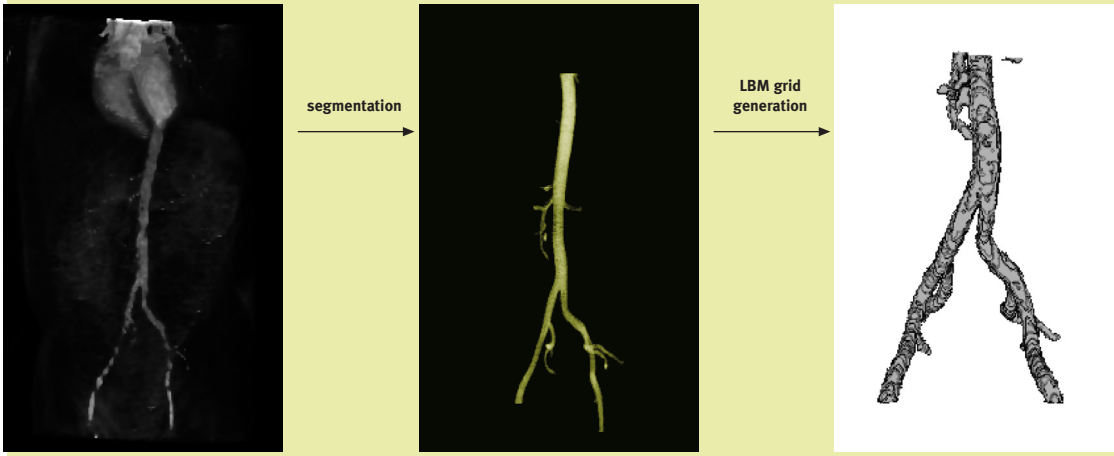


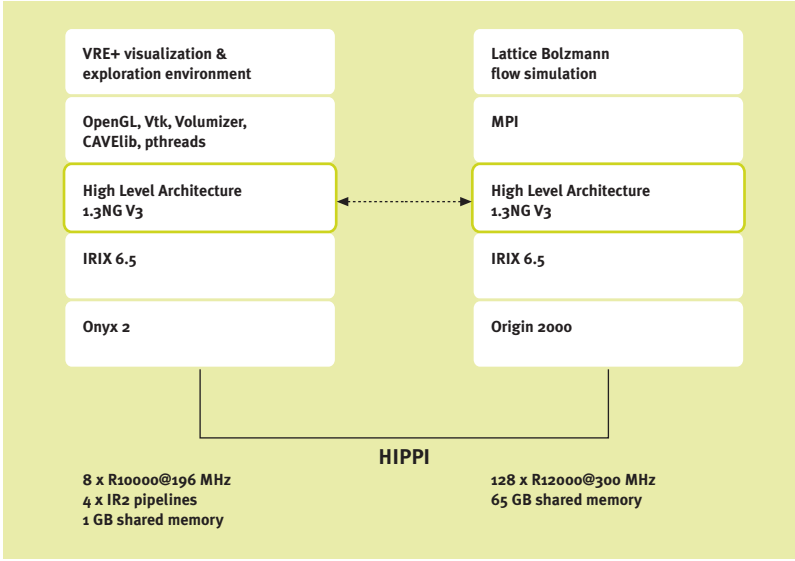
Figure 8
LBM grids are generated from raw medical scans through a combination of segmentation and image processing techniques.

A surgical procedure is simulated through the use of a 3D grid editor. This system allows a user to interactively add and or delete areas in the LBM grid corresponding to the procedure that is simulated. Similar grid generation techniques as described above are used to ensure the grids comply with the demands imposed by LBM.

Middleware

The different components involved in our interactive simulation system are shown in Figure 9. As can be seen from this figure, the visualization and exploration system runs on a different system (a CAVE) than the simulation system (a massively parallel Origin 2000). HLA is used as a middleware layer to connect these components together. By using HLA, the different components can run asynchronously, while spatial and temporal effects as described in the Section on time management can be controlled. In addition, HLA provides attribute ownership management and does efficient data distribution between heterogeneous (if needed) systems.

Figure 9
The visualization and exploration environment (on the left, running in a CAVE) and the simulation system communicate via the HLA.



Discussion and future work

We have presented our views on dynamic exploration environments that support distributed interactive simulation. We have provided an overview on the requirements of such an environment and the issues involved in its construction. We have described how the HLA offers all requirements that are needed in its basic architecture. The case described in the final section is presented as a prototypical case study to validate these assumptions.

Preliminary measurements on the test case environment show that HLA is a suitable architecture to build a relatively efficient interactive and distributed simulation environment. Compared to raw network performance, communication overhead and delays imposed by HLA are acceptable. Implementing a HLA federation, however, requires a substantial effort, but is mostly due to the lack of proper software development tools.

The performance of the test case simulation environment will be validated through a comparison of fluid flow simulation results and the results of other simulation methods as well as in vivo measurements of blood flow through phantom structures and pre- and post-operative MRA scans.

Acknowledgements

This research is funded through grant 612-21-103 from the Netherlands Organization for Scientific Research (NWO). We are also greatly indebted to Charles A. Taylor (Department of Mechanical Engineering, Stanford University) for his insightful and inspiring discussions and for allowing us to use his data sets, Sean A. Spicer (Department of Mechanical Engineering, Stanford

University) for providing his Volumizer Convenience Classes and enabling them for use in the CAVE, Silicon Graphics Inc. for their patience in answering all our questions and fixing bugs in the course of this research, Drona Kandhai (Section Computational Science, The Universiteit van Amsterdam) for his work on the Lattice-Boltzmann fluid simulation environment and Zhiming Zhao (Section Computational Science, The Universiteit van Amsterdam) for his work on HLA and IA's. Finally, we would like to thank Eva Rombouts for her medical input and Alfons Hoekstra for his helpful remarks, while reading this document.

References

- Academic Computing Services Amsterdam (SARA). (1998). Amsterdam, The Netherlands. SARA - CAVE Homepage. http://www.sara.nl/hec/vr/cave/
- Artoli, A.M., D. Kandhai, A.G. Hoekstra, P.M.A. Sloot. (2000). Accuracy of Shear Stress Calculations in the Lattice Boltzmann Method. Accepted for the 9th International Conference on Discrete Simulation of Fluid Dynamics
- Basdogan, C., H. Chih-Hao, M.A. Srinivasan. (1999). Simulation of Tissue Cutting and Bleeding for Laparoscopic Surgery Using Auxiliary Surfaces. In: J.D. Westwood, H.M. Homan, R.A. Robb, D. Stredney. (eds.). (1999). Medicine Meets Virtual Reality. pp38-44. IOS Press, Amsterdam
- Belleman, R.G., J.A. Kaandorp, P.M.A. Sloot. (1998). A Virtual Environment for the Exploration of Diffusion and Flow Phenomena in Complex Geometries. Future Generation Computer Systems **14** (3-4):209-214
- Belleman, R.G., J.A. Kaandorp, D. Dijkman, P.M.A. Sloot. (1999). GEOPROVE: Geometric Probes for Virtual Environments. In: P.M.A. Sloot, M. Bubak, A. Hoekstra, L.O. Hertzberger. (eds.). High Performance Computing and Networking (HPCN'99). pp817-827, Amsterdam, The Netherlands. Springer Verlag
- Bockholt, U., U. Ecke, W. Muller, G. Voss. (1999). Realtime Simulation of Tissue Deformation for the Nasal Endoscopy Simulator (NES). In: J.D. Westwood, H.M. Homan, R.A. Robb, D. Stredney. (eds.). Medicine Meets Virtual Reality. pp74-75. IOS Press, Amsterdam
- Bryson, S. (1996a). Virtual Reality in Scientific Visualization. Communications of the ACM **39** (5):62-71
- Bryson, S., S. Johan. (1996b). Time Management, Simultaneity and Time-Critical Computation in Interactive Unsteady Visualization Environments. Proceedings of Visualization '96. p255. IEEE Computer Science Press, Los Alamitos, CA
- Chen, J.X., D. Rine, H.D. Simon. (1996). Advancing Interactive Visualization and Computational Steering. IEEE Computational Science & Engineering, pp13-17
- Chen, S., G.D. Doolen. (1998). Lattice Boltzmann Method for Fluid Flows. Annu. Rev. Fluid Mech. **30**:329

- Cruz-Neira, C., D.J. Sandin, T.A. DeFanti. (1993). Surround-Screen Projection-Based Virtual Reality: The Design and Implementation of the CAVE. SIGGRAPH '93 Computer Graphics Conference. pp135-142. ACM SIGGRAPH
- Defense Modeling and Simulation Office (DMSO). (1999). Department of Defense, US. High Level Architecture Run Time Infrastructure Programmer's Guide (1.3 version 7). http://hla.dmso.mil/
- Defense Modeling and Simulation Office (DMSO). (1999). High Level Architecture (HLA) homepage. http://hla.dmso.mil/
- Durnford, L. (1999). Virtual Reality: More than just a Game. Radio Netherlands Wereldomroep. http://www.rnw.nl/science/html/virtualreality990514.html
- Elvins, T. (1997). Virtually Lost in Virtual Worlds — Wayfinding without a Cognitive Map. Computer Graphics. http://www.sdsc.edu/~todd/
- Foster, I., C. Kesselman. (1997). Globus: A Metacomputing Infrastructure Toolkit. International Journal Supercomputer Applications **11** (2):115-128
- Hayes-Roth, B. (1995). An Architecture for Adaptive Intelligent Systems. Artificial Intelligence. Special Issue on Agents and Interactivity **72**:329-365
- IBM Corporation, Armonk, NY. (1991). Data Explorer Reference Manual
- Johnson, Ch.R., S.G. Parker. Applications in Computational Medicine Using SCIRun: A Computational Steering Programming Environment. In: H.W. Meuer. (ed.). Supercomputer '95. pp2-19
- Kaandorp, J.A. (ed.). (1998). Future Generation Computer Systems. Special Double Issue on Virtual Reality in Industry and Research **14** (3-4). Elsevier Science
- Kandhai, D., A. Koponen, A.G. Hoekstra, M. Kataja, J. Timonen, P.M.A. Sloot. (1998). Lattice Boltzmann Hydrodynamics on Parallel Systems. Computer Physics Communications
- Kandhai, D. (1999a). Large Scale Lattice-Boltzmann Simulations (Computational Methods and Applications). PhD Thesis. The Universiteit van Amsterdam, Amsterdam, The Netherlands
- Kandhai, D., D. Vidal, A. Hoekstra, H. Hoefsloot, P. Iedema, P.M.A. Sloot. (1999b). Lattice-Boltzmann and Finite Element Simulations of Fluid Flow in a SMRX Mixer. Int. J. Numer. Meth. Fluids **31**:1019-1033
- Koponen, A., D. Kandhai, E. Hellen, M. Alava, A. Hoekstra, M. Kataja, K. Niskanen, P. Sloot, J. Timonen. (1998). Permeability of Three-Dimensional Random Fiber Webs. Physical Review Letters **0** (4):716-719
- Ku, D.N. (1997). Blood Flow in Arteries. Annu. Rev. Fluid Mech. **29**:399-434
- Liere, R. van, J.D. Mulder, J.J. van Wijk. (1996). Computational Steering. In: H. Liddell, A. Colbrook, B. Hertzberger, P. Sloot. (eds.). High-Performance Computing and Networking. pp696-702. Springer Verlag
- Mulder, J.D., J.J. van Wijk. (1995). 3D Computational Steering with Parameterized Geometric Objects. In: G.M. Nielson, D. Silver. (eds.). IEEE

Visualization '95, pp304-312. IEEE CS
- Overeinder, B.J., P.M.A. Sloot. (1999). Extensions to Time Warp Parallel Simulation for Spatially Decomposed Applications. In: D. Al-Dabass, R. Cheng. (eds.). Proceedings of the Fourth United Kingdom Simulation Society Conference (UKSim 99). pp67-73. Cambridge, UK
- Parker, S.G., Ch.R. Johnson. (1995). SCIRun: A Scientific Programming Environment for Computational Steering. Supercomputing '95
- Roy, T.M., C. Cruz-Neira, T.A. DeFanti, D.J. Sandin. (1995). Steering a High Performance Computing Application from a Virtual Environment. Presence: Teleoperators and Virtual Environments **4** (2):121-129
- Schroeder, W., K. Martin, B. Lorensen. (1997). The Visualization Toolkit, an Object-Oriented Approach to 3D Graphics. 2nd edition. Prentice Hall, Upper Saddle River, NJ
- Silicon Graphics Inc. Software Products. (2000). OpenGL homepage. http://www.sgi.com/software/opengl/
- Silicon Graphics Inc. Software Products. (2000). Volumizer homepage. http://www.sgi.com/software/volumizer/
- Surfnet. Gigaport homepage. (2000). http://www.gigaport.nl/en index.html
- Taylor, Ch.A., Th.J.R. Hughes, Ch.K. Zarins. (1998). Finite Element Modeling of Three-Dimensional Pulsatile Flow in the Abdominal Aorta: Relevance to Atherosclerosis. Annals of Biomedical Engineering **26**:975-987
- Taylor, V.E., J. Chen, T.L. Disz, M.E. Papka, R. Stevens. (1996). Interactive Virtual Reality in Simulations: Exploring Lag Time. IEEE Computational Science and Engineering. pp46-54
- The Numerical Algorithms Group Ltd. (1998). Oxford, UK. Iris Explorer User's Guide
- Upson, C., T. Faulhaber, jr., D. Kamins. (et al.). (1989). The Application Visualization System: a Computational Environment for Scientific Visualization. IEEE Computer Graphics and Applications **9** (4):30-42
- Versweyveld, L. (1998). Exploring the Medical Applications of Virtual Reality Techniques in the Dutch CAVE. Virtual Medical Worlds. http://www.hoise.com/vmw/articles/LV-VM-04-98-13.html
- Yucel, E.K., Ch.M. Anderson, R.R. Edelman, Th.M. Grist, R.A. Baum, W.J. Manning, A. Culebras, W. Pearce. (1999). Magnetic Resonance Angiography. Update on Applications for Extracranial Arteries). Circulation **100**:2284-2301

6

6.4 Data Mining Trends

This chapter will discuss several important trends in KDD and data mining. We will start with a view of special hardware for data mining purposes. Building on this, Section 6.4.2 focuses on parallel data mining. Section 6.4.3 discusses mining on relational databases. Meta-learning for automatic technique selection is discussed in Section 6.4.4. Section 6.4.5 focuses on methods to update derived knowledge from changed data.

6.4.1 Computer architectures for data mining

Aad J. van der Steen[1]

Introduction

There are various ways to master the enormous amount of data that is available at present. In scientific computing this flood of data can often be turned into meaningful information by the use of large-scale computational models like weather prediction and climate models or models that describe the structure and interaction of (bio)molecules. Although in such cases huge amounts of data often have to be processed, this usually can be done in a controlled and structured way [Steen, 1995]. By contrast, in data mining we are confronted with a different problem: large amounts of data are available, but neither their structure nor the problem specification are often completely fixed in advance. This calls for facilities, both in hardware and software, that are very flexible and extensible. Such requirements have a direct impact on the computer configurations that can efficiently deal with the problems to be solved. On one hand computer systems should somehow be able to optimize the performance demanded by large-scale data mining applications. On the other hand the hardware and software solutions should be of a sufficient generality to be applicable in a variety of situations.

In this contribution we consider the hardware developments that may help in keeping up with the growing demands that arise from our endeavors in data mining. We will discuss the possibilities and limitations of current hardware and the developments in the near future that may help or hinder the course of this important area in information processing.

Unlike in general scientific computing, the data to be processed will initially always reside on external storage media. The data to be processed will therefore have to be transported to the computer system properly, which may or may not generate intermediate data (which in turn has to be stored on a background medium). In all cases the physical transport of the data will account for a significant part of the total processing time in a data mining application. This means that the requirements for the I/O[2] facilities of the computer system should be in balance with the processing capacity of the system. This is a critical part in the success or failure for the system to cope with the problems it should solve.

It is an important observation that presently no computer system has been developed specifically for data mining. The consequence is that one rather has to make do with the facilities that a system can offer than dictate what the system should be able to do. It is not likely that this situation will change drastically

1 Dr Ir A.J. van der Steen, steen@phys.uu.nl, Computational Physics, Utrecht University, The Netherlands

2 Input/output.

in the next few years. So, it will be the task of the data miner to employ the available computer systems as efficiently as possible, even when their architectures may not be optimal for the work to be done. Fortunately, there are developments in computer and storage architecture that work in favor of data mining, along with other application areas like high-speed communication and multimedia use. The most important development in this respect is parallelism, i.e. the combined operation of many processors working on a common task. Parallelism is already well accepted in the scientific community for large computing-intensive tasks for several years. In data mining parallelism does not (yet) have the same level of acceptance, while the potential of successful application is quite high. In the following sections we will therefore devote a significant part of our discussion to parallel computer systems and the trends that determine their near-term development. In addition, some interesting developments at the processor level will be presented that may have a profound influence on the hardware for data mining.

Parallel systems — The present scene

In computing-intensive scientific disciplines like weather forecasting, airplane design, or quantum-chemical computation parallel processing nowadays is in common use. Indeed, progress in these areas without the use of parallel computers is unthinkable. Also data mining has many characteristics that make it highly suitable for parallel processing, because of the many independent data items that can be processed independently. For very large data mining operations like extracting interesting features from astronomical satellite data the amount of data to be processed can easily involve tens of Terabytes of data (a Terabyte is 10^{12}= a million million bytes). It is evident this can not be done by run-of-the-mill desktop computers and one has to turn to parallel computers to achieve meaningful results.

We will proceed by presenting a common classification of parallel computer systems and by discussing their relevance for data mining.

Parallel system classification

For many years the taxonomy of Flynn [Flynn, 1972] has proven to be useful for the classification of high-performance computers. This classification is based on the way instruction and data streams are manipulated and comprises four main architectural classes. We will first briefly sketch these classes and fill in some details, when each of the classes is described.

– *SISD*[3] *machines*: These are the conventional systems that contain one CPU and hence can accommodate one instruction stream that is executed serially. Nowadays many large mainframes may have more than one CPU, but each of these execute instruction streams that are unrelated. Therefore, such systems still should be regarded as (a couple of) SISD machines acting on dif-

3 Single Instruction stream Single Data stream.

ferent data spaces. The definition of SISD machines is given here for completeness' sake. We will not discuss this type of machine in this report.[4]

- *SIMD*[5] *machines*: Such systems often have a large number of processing units, ranging from 1,024 to 16,384 that all may execute the same instruction on different data in lock-step. So, a single instruction manipulates many data items in parallel.[6]
- *Vector processors* act on arrays of similar data rather than on single data items using specially structured CPUs. When data can be manipulated by these vector units, results can be delivered with a rate of one, two and — in special cases — of three per clock cycle (a clock cycle being defined as the basic internal unit of time for the system). So, vector processors execute on their data in an almost parallel way, but only when executing in vector mode. In this case they are several times faster than in conventional scalar mode. For practical purposes vector processors are therefore mostly regarded as SIMD machines[7].
- *MISD*[8] *machines*: Theoretically in this type of machine multiple instructions should act on a single stream of data. As yet no practical machine in this class has been constructed nor are such systems easily to conceive. We will disregard them in the following discussions.
- *MIMD*[9] *machines*: These machines execute several instruction streams in parallel on different data. The difference from the multi-processor SISD machines mentioned above lies in the fact that the instructions and data are related, because they represent different parts of the same task to be executed. So, MIMD systems may run many subtasks in parallel in order to shorten the time-to-solution for the main task to be executed. There are a large variety of MIMD systems and especially in this class the Flynn taxonomy proves to be not fully adequate for the classification of systems. Systems that behave very differently like a four-processor NEC SX-5 vector system and a thousand processor Cray T3E fall both in this class. In the following we will make another important distinction between classes of systems and treat them accordingly.
- *Shared-memory systems*: Shared-memory systems have multiple CPUs all of which share the same address space. This means that the knowledge of where data is stored is of no concern to the user, as there is only one memory accessed by all CPUs on an equal basis. Shared memory systems can be both SIMD or MIMD. Single-CPU vector processors can be regarded as an example of the former, while the multi-CPU models of these machines are examples of the latter. We will sometimes use the abbreviations SM-SIMD and SM-MIMD for the two subclasses.
- *Distributed-memory systems*: In this case each CPU has its own associated memory. The CPUs are connected by some network and may exchange data between their respective memories, when required. In contrast to shared-

4 Examples of SISD machines are for instance most workstations like those of DEC, Hewlett-Packard, and Sun Microsystems.

5 Single Instruction stream Multiple Data stream.

6 Examples of SIMD machines in this class are the CPP DAP Gamma II and the Quadrics Apemille.

7 An example of such systems is for instance the Hitachi S3600.

8 Multiple Instruction stream Single Data stream.

9 Multiple Instruction stream Multiple Data stream.

memory machines the user must be aware of the location of the data in the local memories and will have to move or distribute these data explicitly when needed. Again, distributed-memory systems may be either SIMD or MIMD. The first class of SIMD systems mentioned, which operate in lock step all have distributed memories associated to the processors. As we will see, distributed-memory MIMD systems exhibit a large variety in the topology of their connecting network. The details of this topology are largely hidden from the user, which is quite helpful with respect to portability of applications. For the distributed-memory systems we will sometimes use DM-SIMD and DM-MIMD to indicate the two subclasses.

As mentioned earlier, although the difference between shared and distributed-memory machines seems clear cut, this is not always entirely the case from the user's point of view. For instance, the late Kendall Square Research systems employed the idea of 'virtual shared-memory' on a hardware level. Virtual shared-memory can also be simulated at the programming level: a specification of High Performance Fortran (HPF) was published in 1993 [High Performance Fortran Forum, 1993] which by means of compiler directives distributes the data over the available processors. Therefore, the system on which HPF is implemented in this case will look like a shared-memory machine to the user. Other vendors of Massively Parallel Processing systems (sometimes called MPP systems), like HP and SGI, also support proprietary virtual shared-memory programming models because these physically distributed memory systems are able to address the whole collective address space. So, for the user such systems have one global address space spanning all of the memory in the system. We will say a little more about the structure of such systems in the section on ccNUMA[10] machines. In addition, packages like TreadMarks [Amza, to appear] provide a virtual shared-memory environment for networks of workstations. Distributed processing takes the DM-MIMD concept one step further: instead of many integrated processors in one or several boxes, workstations, mainframes, etc. are connected by (Gigabit) Ethernet, Fiber Channel, ATM, or otherwise and set to work concurrently on tasks in the same program. Conceptually, this is not different from DM-MIMD computing, but the communication between processors can be much slower.

Packages that initially were made for distributed computing like PVM, Parallel Virtual Machine [Geist, 1994], and MPI, Message Passing Interface [Snir, 1998; Gropp, 1998] have become de facto standards for the 'message passing' programming model. MPI and PVM have become so widely accepted that they have been adopted by all vendors of distributed-memory MIMD systems and even on shared-memory MIMD systems for compatibility reasons. In addition, there is a tendency to cluster shared-memory systems, for instance by HiPPI channels, to

10 Cache Coherent Non Uniform Memory Access.

obtain systems with a very high computational power. E.g. the NEC SX-5, and the Cray SV1 have this structure. So, within the clustered nodes a shared-memory programming style can be used, while between clusters message-passing should be used.
For SM-MIMD systems we should mention OpenMP [Chandra, 2001; OpenMP Forum, 1997] that can be used to parallelize Fortran and C(++) programs by inserting comment directives (Fortran 77/90/95) or pragmas (C/C++) into the code. OpenMP has quickly been adopted by the major vendors and has become a well-established standard for shared memory systems.

Shared-memory SIMD machines

This subclass of machines is practically equivalent to the single-processor vector processors, although other interesting machines in this subclass have existed (viz. VLIW machines [Steen, 1990]). In the block diagram in Figure 1 we depict a generic model of a vector architecture.

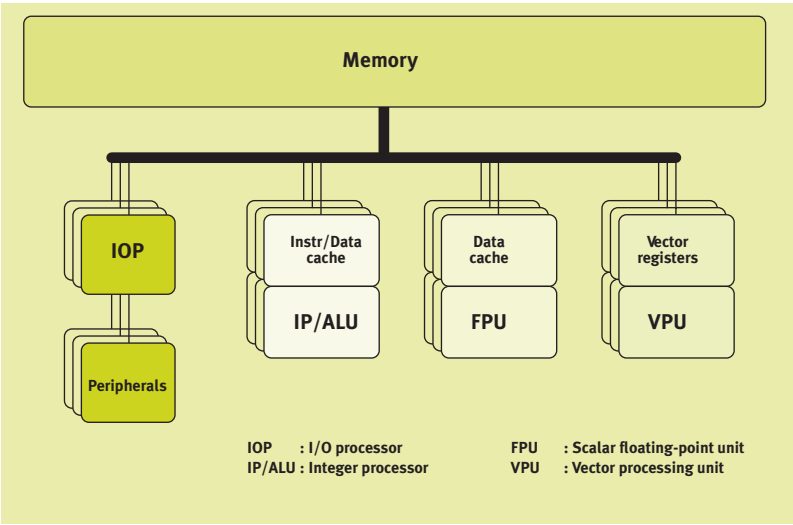


Figure 1
Block diagram of a vector processor.

The single-processor vector machine will have only one of the vector processors depicted and the system may even share its scalar floating-point capability with the vector processor. It may be noted that the Vector Processing Unit (VPU) does not show a cache. The majority of vector processors do not employ a cache anymore. In many cases the vector unit cannot take advantage of it and execution speed may even be unfavorably affected, because of frequent cache overflow.

All present-day vector processors use vector registers to which data are loaded directly. This hardly influences the speed of operations, while providing much more flexibility in gathering operands and manipulation with intermediate results than by loading directly from memory.

Because of the generic nature of Figure 1, no details of the interconnection between the VPU and the memory are shown. Still, these details are very important for the effective speed of a vector operation: when the bandwidth between memory and the VPU is too small, it is not possible to take full advantage of the VPU, because it has to wait for operands and or has to wait before it can store results. When the ratio of arithmetic to load/store operations is not high enough to compensate for such situations, severe performance losses may be incurred. The influence of the number of load/store paths for the dyadic vector operation $c = a + b$ (a, b, and c vectors) is depicted in Figure 2.

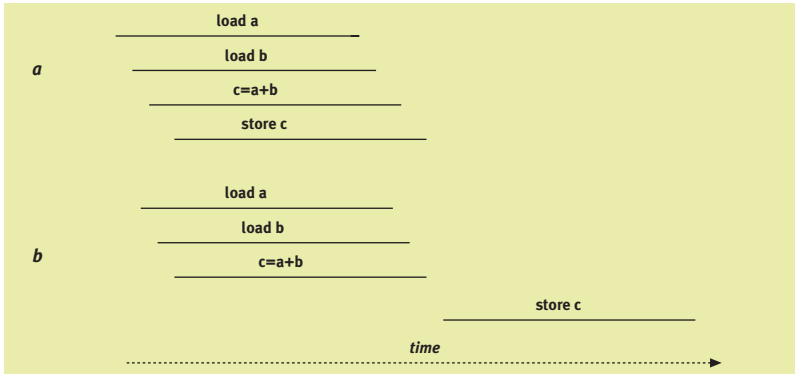


Figure 2
Schematic diagram of a vector addition. Case ***a*** *when two load and one store pipe are available; case* ***b*** *when two load/store pipes are available.*

Because of the high costs of implementing these data paths between memory and the VPU, compromises are often sought and the number of systems that have the full required bandwidth (i.e. two load operations and one store operation at the same time) is limited. In fact, in the vector systems marketed today this high bandwidth does not occur any longer. Vendors prefer to rely on additional caches and other tricks to hide the lack of bandwidth.

The VPUs are shown as a single block in Figure 1. Yet, there is a considerable diversity in the structure of VPUs. Every VPU consists of a number of vector functional units, or 'pipes' that fulfill one or several functions in the VPU. Every VPU will have pipes that are designated to perform memory access functions, thus assuring the timely delivery of operands to the arithmetic pipes and of storing the results in memory again. Usually there will be several arithmetic functional units for integer/logical arithmetic, for floating-point addition, for multiplication and sometimes a combination of both, a so-called compound operation. Division is performed by an iterative procedure, table look-up, or a combination of both using the add and multiply pipe. In addition, there will almost always be a mask pipe to enable operation on a selected subset of elements in a vector of operands. Lastly, such sets of vector pipes can be replicated within one VPU (2 up to 16-fold replication occurs). Ideally, this will increase the performance per VPU by the same factor provided the bandwidth to memory is adequate.

Suitability of vector processors for data mining

At first glance, vector processors have characteristics that make them rather unsuitable for data mining. Traditionally, these machines have always been designed to optimize the throughput of floating-point operations which has less of a priority for the majority of data mining algorithms. Furthermore, algorithms that contain a large proportion of conditional code are known to perform poorly on vector systems, while virtually all data mining algorithms contain a significant amount of conditional code. Lastly, vector processors are all of a proprietary nature, which makes them costly in comparison to the standard processors of which billions may be produced.

Still, there are also some arguments in favor of vector processors: first, the very high bandwidth from memory to the processors without degradation by cache misses. Second, the low instruction count per cycle: when an operation executes in a VPU, we have only one instruction per (set of) vector register(s), thus saving instruction load and decode time. The disadvantage of conditional code may be mitigated by the use of the mask registers that enable all elements in a vector that should execute to be singled out, when the testing condition is true, while disregarding the other elements. In addition, some vector processors contain bit-level matrix multiply units that have been demonstrated to be extremely efficient in decoding operations and have recently also been demonstrated to be orders of magnitude faster in certain protein matching operations than common (RISC) processors. With suitable modification, there is no reason that such systems could not be applied in other pattern matching algorithms. In the present situation, it may turn out that vector processors could be quite cost-effective in a subset of data mining algorithms, the more so as traditionally much attention have been given to the design of fast I/O subsystems on vector processors. In a later section we will consider possibilities of customizing processors relative to the tasks to be fulfilled. Here vector processor-like behavior may be a definite asset.

Distributed-memory SIMD machines

Machines of this type are sometimes also known as processor-array machines [Hockney, 1987]. Because the processors of these machines operate in lock-step, i.e. all processors execute the same instruction at the same time (but on different data items), no synchronization between processors is required. This greatly simplifies the design of such systems. A control processor issues the instructions that are to be executed by the processors in the processor array. All currently available DM-SIMD machines use a front-end processor to which they are connected by a data path to the control processor. Operations that cannot be executed by the processor array or by the control processor are offloaded to the front-end system. For instance, I/O may be through the front-end system, by the processor array machine itself or both. Figure 3 shows a generic model of a

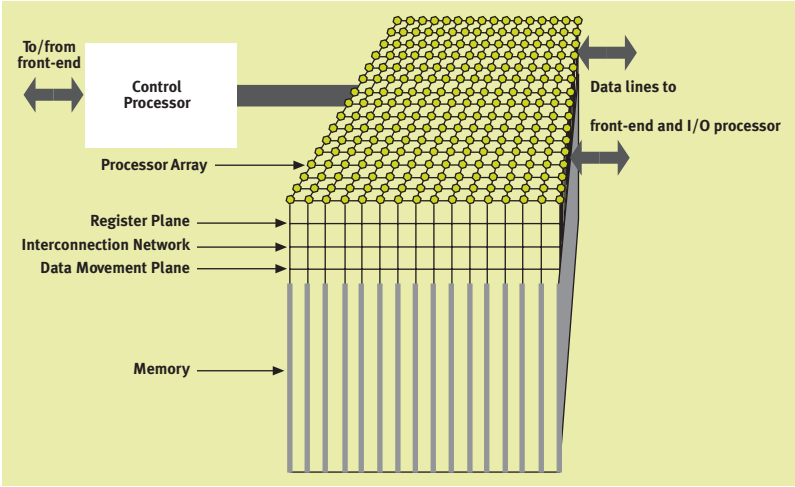


Figure 3
A generic block diagram of a distributed-memory SIMD machine.

DM-SIMD machine of which actual models will deviate to some degree. Figure 3 might suggest that all processors in such systems are connected in a 2D grid and, indeed, the interconnection topology of this type of machine always includes the 2D grid. As opposing ends of each grid line are also always connected the topology is rather that of a torus. This is not the only interconnection scheme: they might also be connected in 3D, diagonally, or more complex structures.

It is possible to exclude processors in the array from executing an instruction on certain logical conditions, but this means that for the time of this instruction these processors are idle (a direct consequence of the SIMD-type operation), which immediately lowers the performance. Another factor that may adversely affect the speed occurs, when data required by processor *i* resides in the memory of processor *j* — in fact, as this occurs for all processors at the same time, this effectively means that data will have to be permuted across the processors. To access the data in processor *j*, the data will have to be fetched by this processor and then send through the routing network to processor *i*. This may be fairly time-consuming. For both reasons mentioned DM-SIMD machines are rather specialized in their use, when one wants to employ their full parallelism. Generally, they perform excellently on digital signal and image processing and on certain types of Monte Carlo simulations[11] where virtually no data exchange between processors is required and exactly the same type of operations is done on massive datasets with a size that can be made to fit comfortably in these machines.

The control processor as depicted in Figure 3 may be more or less intelligent. It issues the instruction sequence that will be executed by the processor array. In the worst case (that means a less autonomous control processor), when an instruction is not fit for execution on the processor array (e.g. a simple print

11 Method making use of random numbers and probability statistics. See [Sobol, 1994].

instruction), it might be offloaded to the front-end processor, which may be much slower than execution on the control processor. In case of a more autonomous control processor this can be avoided, thus saving processing interrupts both on the front-end and the control processor. Most DM-SIMD systems have the possibility to handle I/O independently from the front/end processors. This is not only favorable, because the communication between the front-end and back-end systems is avoided. The (specialized) I/O devices for the processor-array system are generally much more efficient in providing the necessary data directly to the memory of the processor array. Especially for very data-intensive applications like radar and image processing such I/O systems are very important.

A feature that is peculiar to this type of machines is that the processors are sometimes of a very simple bit-serial type, i.e. the processors operate on the data items bitwise, irrespective of their type. Consequently operations on integers are produced by software routines on these simple bit-serial processors, which takes at least as many cycles as the operands are long. So, a 32-bit integer result will be produced two times faster than a 64-bit result. For floating-point operations a similar situation holds, be it that the number of cycles required is a multiple of that needed for an integer operation. As the number of processors in this type of systems is mostly large[12], the slower operation on floating-point numbers can be often compensated for by their number, while the cost per processor is quite low as compared to full floating-point processors. In some cases, however, floating-point coprocessors are added to the processor-array. Their number is 8-16 times lower than that of the bit-serial processors, because of the cost argument. An advantage of bit-serial processors is that they may operate on operands of any length. This is particularly advantageous for random number generation (which often boils down to logical manipulation of bits) and for signal processing, because in both cases operands of only 1-8 bits are abundant. As the execution time for bit-serial machines is proportional to the length of the operands, this may result in significant speedups.

Suitability of processor-array systems for data mining

Processor-array machines were initially designed for signal and image processing. In addition to a highly compute-intensive part in image processing there has always been a less compute-intensive part concerned with feature extraction. Processor-array systems are very well suited for this kind of task, because each image point is largely independent from the surrounding points and they can manipulate data items that are not necessarily standard data types. Therefore, processor-arrays can be (and frequently are) used in full text searches, satellite image differencing and matching, and recently in bioinformatics applications.

12 1024 Or larger, the Quadrics Apemille is a notable exception, however.

Although processor-array systems have many characteristics that are conductive to a subset of data mining applications, it is not likely that these systems will play a major role in the near future. This has several reasons. One of the arguments of designing processor-array systems was the relatively low complexity of such systems as compared to MIMD-type machines. Of course this low complexity should show in the development cost and the cost-effectiveness of the systems. Unfortunately, this is only partly the case, because the network connecting the processors is a cost factor that is at least as important as the processor part of the machine. As general processors have become relatively cheap, the competitiveness of the simple array processors has significantly decreased over the last few years. When one considers that special-purpose software is also needed to take advantage of the architecture, it will be clear that it is very hard to compete with general-purpose parallel systems for which the software has a much better perspective of portability. For these reasons, processor-array machines will probably disappear from the parallel systems scene, except may be for specialized defense purposes.

Shared-memory MIMD machines

In Figure 1 already one subclass of this type of machine was shown. In fact, the single-processor vector machine discussed there was a special case of a more general type. The figure shows that more than one FPU[13] and or VPU may be possible in one system.

The main problem one is confronted with in shared-memory systems is that of the connection of the CPUs to each other and to the memory. As more CPUs are added, the collective bandwidth to the memory ideally should increase linearly

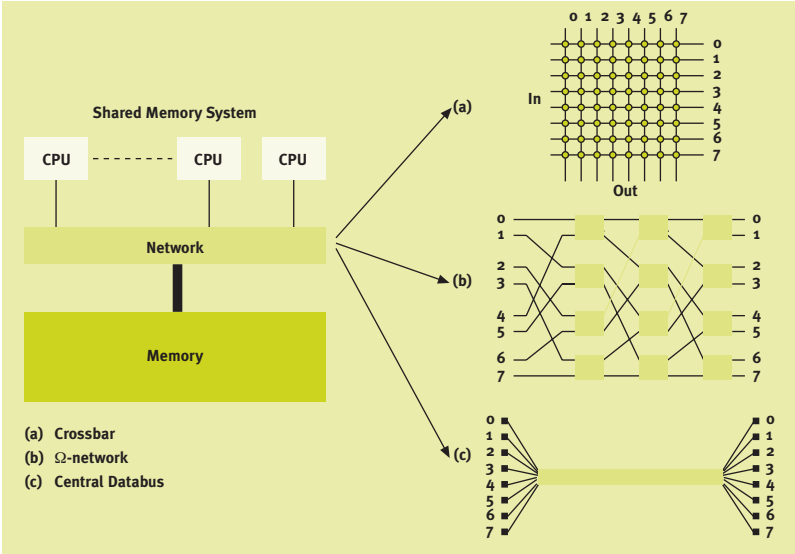


Figure 4
Some examples of interconnection structures used in shared-memory MIMD systems.

13 Floating-Point Unit.

with the number of processors, while each processor should preferably communicate directly with all others without the much slower alternative of having to use the memory in an intermediate stage. Unfortunately, full interconnection is quite costly, growing with $O(n^2)$, while increasing the number of processors with $O(n)$. So, various alternatives have been tried. Figure 4 shows some of the interconnection structures that are (and have been) used.

As can be seen from Figure 4, a crossbar uses n^2 connections, an Ω-network uses $n\log_2 n$ connections, while with the central bus there is only one connection. This is reflected in the use of each connection path for the different types of interconnections: for a crossbar each data path is direct and does not have to be shared with other elements. In case of the Ω-network there are $\log_2 n$ switching stages and as many data items may have to compete for any path. For the central data bus all data have to share the same bus, so n data items may compete at any time.

The bus connection is the least expensive solution, but it has the obvious drawback that bus contention may occur, thus slowing down the computations. Various intricate strategies have been devised using caches associated with the CPUs to minimize the bus traffic. This leads, however, to a more complicated bus structure which raises the costs. In practice it has proved to be very hard to design buses that are fast enough, especially where the speed of the processors has been increasing very quickly and it imposes an upper bound on the number of processors thus connected that in practice appears not to exceed a number of 10-20. In 1992, a new standard (IEEE P896) for a fast bus to connect either internal system components or to external systems has been defined. This bus, called the Scalable Coherent Interface (SCI) should provide a point-to-point bandwidth of 200-1,000 MB/s. It is in fact used in some computer systems, but also within clusters of workstations. The SCI is much more than a simple bus and it can act as the hardware network framework for distributed computing [James, 1990].
A multi-stage crossbar is a network with a logarithmic complexity and it has a structure, which is situated somewhere in between a bus and a crossbar with respect to potential capacity and costs. The Ω-network as depicted in Figure 4 is an example[14]. It is quite conceivable that new machines may use it, especially as the number of processors grows. For a large number of processors the $n\log_2 n$ connections quickly become more attractive than the n^2 used in crossbars. Of course, the switches at the intermediate levels should be sufficiently fast to cope with the bandwidth required. Obviously, not only the structure, but also the width of the links between the processors is important: a network using 16-bit parallel links will have a bandwidth, which is 16 times higher than a network with the same topology implemented with serial links.

14 Commercially available machines like the IBM RS/6000 SP, the SGI Origin3000, and the NEC Cenju-4 use such a network structure, next to a number of experimental machines.

In all present-day multi-processor vector processors crossbars are used. This is still feasible because the maximum number of processors in a system is still rather small (32 at most presently). When the number of processors would increase, however, technological problems might arise. Not only does it become harder to build a crossbar of sufficient speed for the larger numbers of processors, the processors themselves generally also increase in speed individually, doubling the problems of making the speed of the crossbar match that of the bandwidth required by the processors.

Whichever network is used, the type of processors in principle could be arbitrary for any topology. In practice, however, bus structured machines do not have vector processors as the speeds of these would grossly mismatch with any bus that could be constructed with reasonable costs. All available bus-oriented systems use RISC processors. The local caches of the processors can sometimes alleviate the bandwidth problem, if the data access can be satisfied by the caches, thus avoiding references to the memory.

The systems discussed in this subsection are of the MIMD type and therefore different tasks may run on different processors simultaneously. In many cases synchronization between tasks is required and again the interconnection structure is very important here. Most vector processors employ special communication registers within the CPUs by which they can communicate directly with the other CPUs they have to synchronize with. The systems may also synchronize via the shared memory. Generally, this is much slower, but it can still be acceptable, when the synchronization is relatively seldom. Of course, in bus-based systems communication also has to be done via a bus. This bus is mostly separated from the data bus to ensure a maximum speed for the synchronization.

Suitability of SM-MIMD systems for data mining

In principle shared-memory MIMD systems should be quite capable of solving data mining problems of all kinds. Interestingly, surveys like [Zaki, 1999] show that research in parallelizing for instance association rule mining on shared-memory systems and also the actual use are less intense than on distributed-memory systems. This may at least partly be explained by a lack of standardization for shared-memory programming. The introduction of OpenMP [OpenMP Forum, 1997], now the de facto standard, has markedly improved this situation. However, OpenMP implementations are still relatively immature sometimes leading to system overheads that result in lower speed-up of applications than could be expected.

The non-vector processor based SM-MIMD systems, also known as Symmetric Multi Processor (SMP) systems presently have a number of processors that is limited to about 32-64, because of the inherent bandwidth limitations for larger

systems (but see also the section on ccNUMA machines). The main problem in these systems is to keep as much of the data and instructions in the caches as the bandwidth from the memory to the processors is insufficient to feed the processors. All SMP systems are cache-based and unfortunately almost no vendor provides Operating System support to place the data optimally onto the processors. As data mining algorithms are inherently data-hungry, the limiting factor will in fact be the collective memory bandwidth and not the processor speed.
Nevertheless, the relative ease of modifying sequential data mining algorithms for use in a SM-MIMD environment makes this class of machines attractive and one may expect that there will a significant growth in parallel algorithms for these systems in the next few years.

Distributed-memory MIMD machines

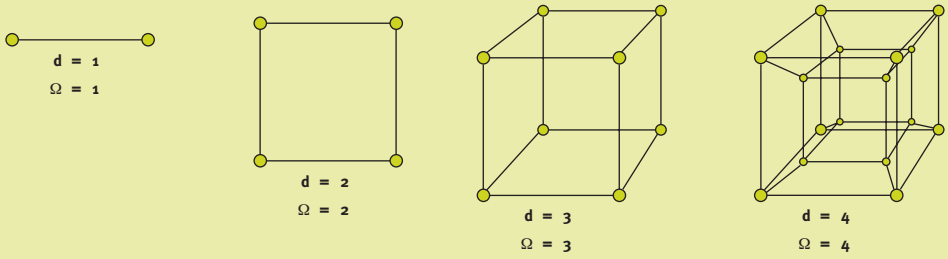
The class of DM-MIMD machines is the fastest growing part in the family of high-performance computers. Although this type of machine is more difficult to deal with than shared-memory machines and DM-SIMD machines. The latter types of machine are processor-array systems in which the data structures that are candidates for parallelization are vectors and multidimensional arrays that are laid out automatically on the processor array by the system software. For shared-memory systems the data distribution is completely transparent to the user. This is quite different for DM-MIMD systems where the user has to distribute the data over the processors and where the data exchange between processors also has to be performed explicitly. The initial reluctance to use DM-MIMD machines seems to have decreased. Partly, this is due to the now existing standard for communication software [Geist, 1994; Snir, 1998; Gropp, 1998] and partly because, at least theoretically, this class of systems is able to outperform all other types of machines.

The advantages of DM-MIMD systems are clear: the bandwidth problem that haunts shared-memory systems is avoided, because the bandwidth scales up automatically with the number of processors. Furthermore, the speed of the memory, which is another critical issue with shared-memory systems (to get a peak performance that is comparable to that of DM-MIMD systems, the processors of the shared-memory machines should be very fast and the speed of the memory should match it) is less important for the DM-MIMD machines, because more processors can be configured without the aforementioned bandwidth problems.

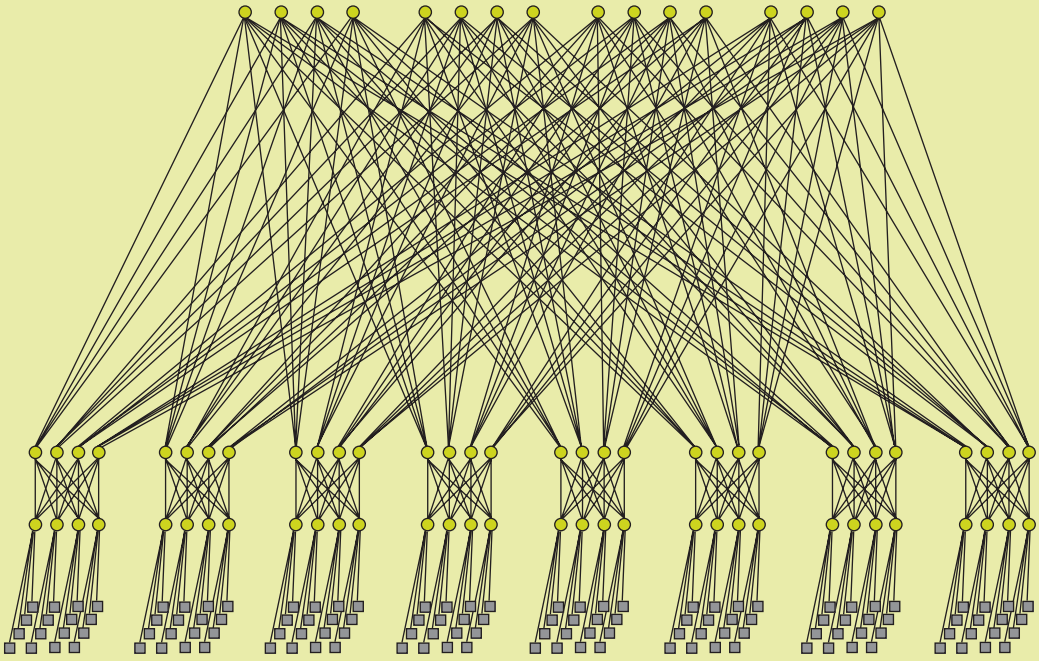
Of course, DM-MIMD systems also have their disadvantages: the communication between processors is much slower than in SM-MIMD systems, and so, the synchronization overhead in case of communicating tasks is generally orders of

magnitude higher than in shared-memory machines. Moreover, the access to data that are not in the local memory belonging to a particular processor has to be obtained from non-local memory (or memories). This is again very slow on most systems as compared to local data access. When the structure of a problem dictates a frequent exchange of data between processors and or requires many processor synchronizations, it may well be that only a very small fraction of the theoretical peak speed can be obtained. As already mentioned, the data and task decomposition are factors that mostly have to be dealt with explicitly, which may be far from trivial.

It will be clear from the paragraph above that also for DM-MIMD machines both the topology and the speed of the data paths are of crucial importance for the



(a) Hypercubes, dimension 1-4.



(b) A 128-way fat tree.

Figure 5
Some often used networks for DM machine types.

practical usefulness of a system. Again, as in the section on SM-MIMD systems, the richness of the connection structure has to be balanced against the costs. Of the many conceivable interconnection structures only a few are popular in practice. One of these is the so-called hypercube topology as depicted in Figure 5(a). A nice feature of the hypercube topology is that for a hypercube with 2^d nodes the number of steps to be taken between any two nodes is at most d. So, the dimension of the network grows only logarithmically with the number of nodes. In addition, theoretically, it is possible to simulate any other topology on a hypercube: trees, rings, 2D and 3D meshes, etc. In practice, the exact topology for hypercubes does not matter too much anymore, because all systems in the market today employ what is called 'worm hole routing'. This means that when a message is send from node *i* to node *j*, a header message is sent from *i* to *j*, resulting in a direct connection between these nodes. As soon as this connection is established, the data proper is sent through this connection without disturbing the operation of the intermediate nodes. Except for a small amount of time in setting up the connection between nodes, the communication time has become virtually independent of the distance between the nodes. Of course, when several messages in a busy network have to compete for the same paths, waiting times are incurred as in any network that does not directly connect any processor to all others, and often re-routing strategies are employed to circumvent busy links.

Another cost-effective way to connect a large number of processors is by means of a fat tree. In principle a simple tree structure for a network is sufficient to connect all nodes in a computer system. However, in practice it turns out that congestion occurs near the root of the tree, because of the concentration of messages that first have to traverse the higher levels in the tree structure, before they can descend again to their target nodes. The fat tree amends this shortcoming by providing more bandwidth (mostly in the form of multiple connections) in the higher levels of the tree. An example of a fat tree with a bandwidth in the highest level that is doubled with respect to the lower levels is shown in Figure 5b.

A number of massively parallel DM-MIMD systems seem to favor a 2 or 3D mesh (torus) structure. The rationale for this seems to be that most large-scale physical simulations can be mapped efficiently on this topology and that a richer interconnection structure hardly pays off. However, some systems maintain (an) additional network(s) besides the mesh to handle certain bottlenecks in data distribution and retrieval [Horie, 1991].

A large fraction of systems in the DM-MIMD class employ crossbars. For relatively small amounts of processors (in the order of 64) this may be a direct or 1-stage crossbar, while to connect larger numbers of nodes multi-stage crossbars are used, i.e. the connections of a crossbar at level 1 connect to a crossbar at level 2, etc., instead of directly to nodes at more remote distances in the topolo-

gy. In this way it is possible to connect in the order of a few thousands of nodes through only a few switching stages. In addition to the hypercube structure, other logarithmic complexity networks like Butterfly, Ω, or shuffle-exchange networks are often employed in such systems.
As with SM-MIMD machines, a node may in principle consist of any type of processor (scalar or vector) for computation or transaction processing together with local memory (with or without cache) and, in almost all cases, a separate communication processor with links to connect the node to its neighbors. Nowadays, the node processors are mostly off-the-shelf RISC processors sometimes enhanced by vector processors. A problem that is peculiar to this DM-MIMD systems is the mismatch of communication versus computation speed that may occur, when the node processors are upgraded without also speeding up the intercommunication. In some cases this may result in turning computational-bound problems into communication-bound problems.

Suitability of DM-MIMD systems for data mining

As will be clear from the beginning of this section, the bandwidth problem that is a limiting factor for SM-MIMD systems is less of a problem in distributed-memory systems. So, in theory, DM-MIMD systems should be very well suited to support parallel data mining. Indeed, the research in parallel data mining has been most active for this class of machines [Zaki, 1999]. This is despite the fact that the development of parallel algorithms for DM-MIMD systems is by no means trivial: not only has the user to decide how to distribute his or her data over the processors, one also has to be very careful in setting up the communication structure for the algorithm at hand, because the time spent in communication strongly affects the scalability of the algorithm and thus its success or failure for a large number of processors. Modest successes have been reported for some association rule mining algorithms ([Han, 1997; Cheung, 1996]). For other, more compute-intensive algorithms the results are relatively better, because the computation/communication ratio is more favorable.
Also, there is a tendency not to work fully in parallel, but to employ replication. In this case data are distributed over the processors and the partial data sets are processed. However, the communication steps that would be required to obtain global results are omitted and approximate solutions from the separate processors are accepted as an approximate solution. In many cases this may work well for a limited number of processors. However, when executed on a larger number of processors, e.g. > 100, this fragmentation may not be useful anymore.
Future development will most probably occur on a class of machines that has grown explosively in the last few years and that will largely dominate the scene of high-performance computers for many years to come. This class of system is discussed in the following section.

ccNUMA machines

As already mentioned in the introduction, a trend can be observed towards building systems that have a rather small (up to 16) number of RISC processors that are tightly integrated in a cluster, a Symmetric Multi-Processing (SMP) node. The processors in such a node are virtually always connected by a 1-stage crossbar, while these clusters are connected by a less costly network. Such a system may look as depicted in Figure 6 . Note that in Figure 6 all CPUs in a cluster are connected to a common part of the memory.

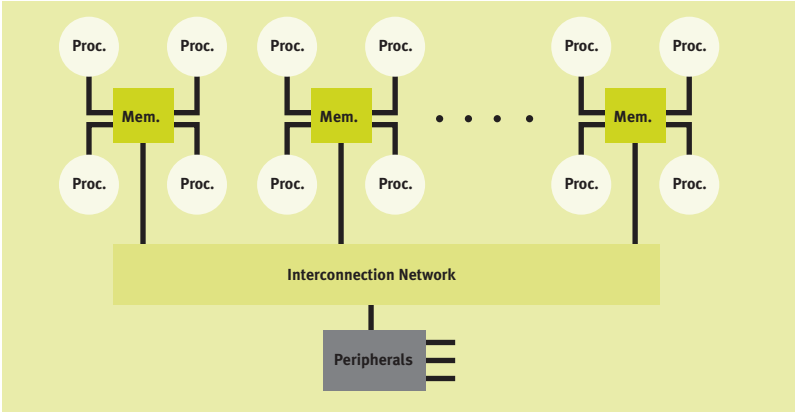


Figure 6
Block diagram of a system with a 'hybrid' network: clusters of four CPUs are connected by a crossbar. The clusters are connected by a less expensive network, e.g. a Butterfly network.

This is similar to the policy mentioned for large vector processor ensembles mentioned above, but with the important difference that all of the processors can access all of the address space, if necessary. The most important ways to let the SMP nodes share their memory are S-COMA (Simple Cache-Only Memory Architecture) and ccNUMA, which stands for Cache Coherent Non-Uniform Memory Access. Therefore, such systems can be considered as SM-MIMD machines. On the other hand, because the memory is physically distributed, it cannot be guaranteed that a data access operation will always be satisfied within the same time. In S-COMA systems the cache hierarchy of the local nodes is extended to the memory of the other nodes. So, when data is required that does not reside in the local node's memory, it is retrieved from the memory of the node where it is stored. In ccNUMA this concept is further extended in that all memory in the system is regarded (and addressed) globally. So, a data item may not be physically local, but belong logically to one shared address space. Because the data can be physically dispersed over many nodes, the access time for different data items may well be different, which explains the term non-uniform data access. The term 'Cache Coherent' refers to the fact that for all CPUs any variable that is to be used must have a consistent value. Therefore, it must be ensured that the caches that provide these variables are also consistent in this respect. There are various ways to ensure that the caches of the CPUs are coherent. One is the snoopy bus protocol in which the caches listen in on trans-

port of variables to any of the CPUs and update their own copies of these variables, if they have them. Another way is the directory memory, a special part of memory which enables to keep track of all the copies of variables and of their validity.
Presently, no commercially available machine uses the S-COMA scheme. By contrast, there are several popular ccNUMA systems[15] commercially available.
For all practical purposes we can classify these systems as being SM-MIMD machines, also because special assisting hardware/software (such as a directory memory) has been incorporated to establish a single system image, although the memory is physically distributed.

Suitability of ccNUMA systems for data mining

In economical terms ccNUMA systems are a highly successful design, which, as a bonus, can be regarded by the user as a shared-memory system and therefore greatly eases the programming effort to be invested to parallelize the data mining software at hand. There is of course the matter of the non-uniformity of data access, but with proper optimization the majority of instruction fetches and data accesses will occur from the respective caches and will not cause intolerable asynchronicity. The trend for the near future will be that more ccNUMA systems will appear in the marketplace and, with it, the software on the operating system and parallel file system level. For the latter kind of software at present no universally accepted standards are available yet, but the research in this field is very active and one even might expect that distributed memory I/O libraries like MPI-IO [Gropp, 1998] or a portable version of GPFS [Barkes, 1998] will be leading the way to optimized parallel I/O for ccNUMA systems also.
Although there is much that speaks for ccNUMA systems, at present there certainly are also still drawbacks. These pertain mainly to the performance of the operation system, the handling of parallel threads, synchronization, and in keeping the caches of the processors coherent. Inefficiencies for all of these issues tend to slow down the system, especially for large numbers of processors. In fact, these are the same performance issues that haunt some SMP systems, but with the extra complicating factor of global synchronization and cache coherence in the system.
The performance losses mentioned above will in many cases not offset the advantages of a shared-memory programming model, especially for a modest number, say of 16-32 processors. With the improvement of OS and I/O functions to be expected for ccNUMA systems, one therefore can expect these systems to play an important role in data mining in the near future.

Future developments

In the previous sections we have discussed the high-performance computer landscape as it exists today and which determines the computers that will reach

15 HP SuperDome, SGI Origin3000.

the market in the next few years. However, there are a number of recent developments that will strongly influence computer architectures in a slightly longer term, say 5 years from now of which the trends are already discernible. Most of these are related to the core of the computer systems, the processor and memory structure, while the developments in high-density storage also will have a high impact on the possibilities for large-scale data mining.

Beowulf clusters

The adoption of clusters, collections of workstations/PCs connected by a local network, has virtually exploded, since the introduction of the first Beowulf cluster in 1994. The attraction lies in the (potentially) low cost of both hardware and software and the control that builders/users have over their system. The interest in clusters can be seen for instance from the active IEEE Task Force on Cluster Computing (TFCC), which regularly issues a White Paper in which the current status of cluster computing is reviewed [Baker, 2000]. Books how to build and maintain clusters have greatly added to their popularity [Sterling, 1999; Spector, 2000]. As the cluster scene becomes relatively mature and an attractive market, large HPC vendors as well as many start-up companies have entered the field and offer more or less ready to wear cluster solutions for those groups that do not want to build their cluster from scratch. Presently, they have reached an appreciable level of maturity, as is testified by a throughput environment study [Steen, 2000].

Beowulf clusters have not gone unnoticed in the data mining community, but it is without doubt that they will play a much larger role in the near future. First, in many respects the performance of clusters is at a comparable level to that of integrated parallel machines, especially where it concerns integer and logical operations. Because a large proportion of operations in data mining are of this kind, the performance of clusters is relatively favorable for this field. Second, the networks that are available for connecting the processors in the clusters are attaining speeds that also are becoming comparable to those in integrated parallel machines. Furthermore, relatively cheap large local disks can be incorporated in the parallel nodes, which can be helpful in the cleaning and filtering stages of raw data. Lastly, the good price/performance ratio of clusters favors the installation of dedicated systems, which in turn enables the optimization of configurations for the data mining task(s) they were meant for.

Processor developments

Until recently processor developments were, within economical constraints, almost exclusively determined by their floating-point performance. Even for markets where floating-point operations did not constitute the most important part of the work, the processors offered did not leave a choice in this respect. This attitude is changing rather quickly under the influence of large application

areas that will absorb the vast majority of processors in the coming years: communication and embedded processors of all sorts. In these areas floating-point processing has a lower priority. Rather, low energy consumption and (for communication processors) fast integer arithmetic are more important. As more and more tasks are amenable to (cooperative) processing and the resources of processor developing teams are finite, one can predict that a diversification in processor types will occur in the next 5 to 10 years.

An important factor in this diversification is the availability of Field Programmable Gate Arrays (FPGAs). FPGAs can be regarded as chips with a collection of 'blank' gates in which the user can define its own circuits to execute exactly the operations desired directly in hardware. As the device density in FPGAs is high and no compromises with regard to other functions on the chip have to be made, this 'hardware programming' can in some cases be orders of magnitude faster than the equivalent operations from a general-purpose processor. FPGAs are currently used in specialized signal processing and (de)coding processors. Because FPGAs are reconfigurable, one is able to switch between various specialized tasks, be it only 100-1,000 times per second. For dedicated systems this need not be a problem and one can imagine that such systems could be configured for searching algorithms, tree building, etc. In fact, the first machines combining standard processors and FPGAs will reach the market within a year. In Figure 7 we show a block diagram of a system that will be commercialized soon, the SRC-6.

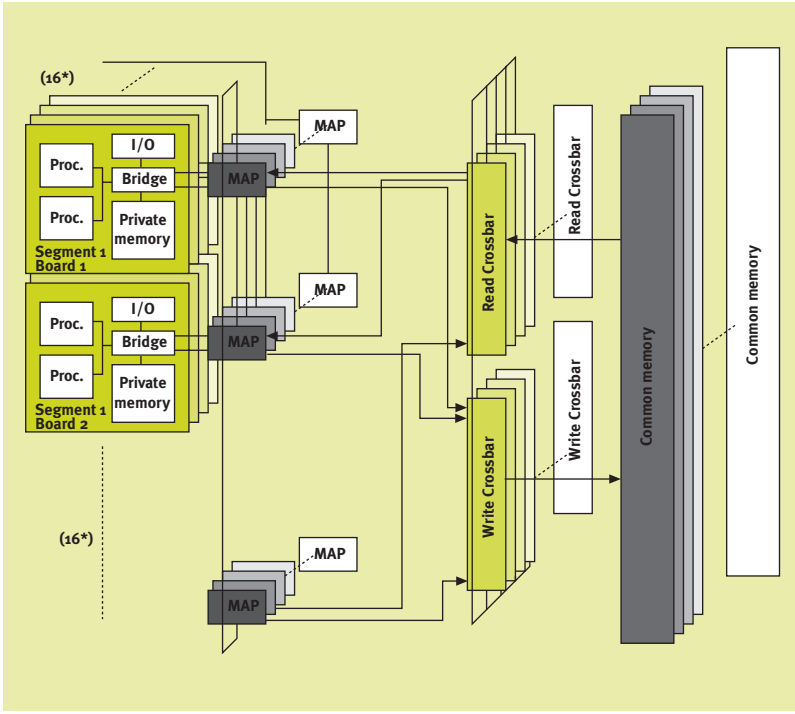


Figure 7
Block diagram of the SRC-6 system that incorporates both conventional processor and FPGAs in its MAP processors.

Each MAP (Multi-Adaptive Processor) in Figure 7 contains a so-called User Logic block, the FPGA proper, a local memory, used by the User Logic block, the MAP control that defines the working of the User Logic block once it has been configured, and configuration ROMs that enables the configuration of the User Logic for a desired function.
A full system consists of 256 nodes. Each node contains two 'normal' (Intel) processors, a private memory, and a local I/O capability. Each node is connected to a MAP and through it to a read and a write crossbar to a common memory, while the MAPs are connected by a double ring network. Depending on the functions programmed into the MAPs, the system could act as a variety of special-purpose systems with considerable power. The potential this could have for dedicated data mining machines is clear: such a system could be configured as a searching machine, a clustering machine, or whatever function is required, which could greatly speed up the total data mining process.

High density storage

With respect to CPU speed Moore's Law, which states that processor speed doubles every 18 months, holds already for a surprising 15 years. The expectation is that this 'law' will be followed for at least another 5 to 7 years, although it becomes increasingly difficult to keep pace. For other system components there are similar growth laws: the speed of memory growth markedly slower than that of CPUs (about a doubling in speed every 2 years), while storage on magnetic disks since 1991 with the introduction of magneto-resistive read/write heads the disk I/O speed doubles every 12 months. This speed growth stems from both a higher revolution speed of the disks and a higher storage density on the disk's surface. The latter part is the most important, which explains that the bandwidth, i.e. the amount of MB/s to be read/written has improved much more than the latency, the time it takes before the first byte actually can be transported. The phenomenal growth in disks speeds in the last 10 years has had the curious by-effect that research in some promising alternative storage ideas have met with less interest than they deserve, because of their high potential or the special features inherent in these new approaches.
Of course the raw speed of a storage medium is not enough to be useful in actual computer systems. One also needs protocols and connection networks that match the speeds that can be delivered by the storage hardware. Presently, on individual machines bandwidths of about 100 MB/s can typically be attained with, e.g. Fiber Channel and 800 MB/s with GSN (Gigabit System Network), in a range of 250 meters using optical cables. Recently, a consortium of major computer and network vendors have completed the definition of the Infiniband protocol, which defines protocols for links that can move data at 250 MB/s per link with 1, 4, or 12 links in parallel. It is expected that the first Infiniband products will reach the market in 2002. With bandwidths in this speed range the distinc-

tion between internal and external disks virtually disappears and this can have a large impact on how the user goes about storing and retrieving data. Obviously data mining will also be affected by this development.

We will conclude by discussing a storage research item that could become of considerable significance for data storage and, a fortiori for data mining.

Holographic storage

Since about 1990 there has been an effort to store information in a solid medium by means of holographic techniques. A laser beam is split into a data bearing beam and a reference beam. The data bearing beam is fed the data to be stored through a Spatial Light Modulator (SLM), which encodes the data. By a system of lenses the data and reference beams shine into a photo-refractive medium like $LiNbO_3$, Lithium Niobiate, or more recently a photo-polymer compound. The data is thus imprinted throughout the medium as an interference wave pattern or 'page'. By slightly altering the angle of the incident reference beam, one can store large amounts of data in the same medium without interference of the different pages. In addition, by using different wavelengths other unique wave patterns can be stored in the medium, multiplying the storage capacity. The data can be read out by shining an unmodulated beam through the medium and detecting the intensity of the signal beam with an array of Charge Coupled Devices (CCDs). The beam steering device that determines the angle at which the reference beam enters the storage medium is also used for addressing, as there is a one-to-one correspondence between the incident angle and the wave pattern that represents the data. The procedure is schematically laid out in Figure 8.

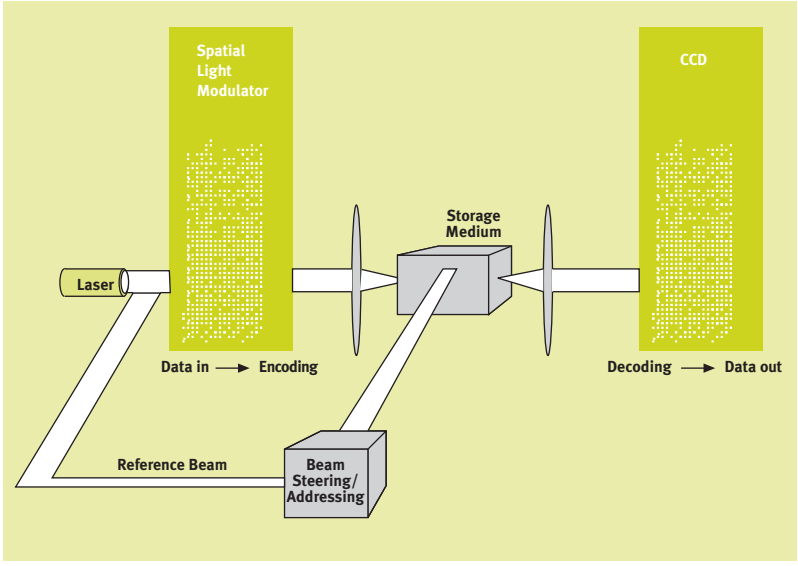


Figure 8
Schematic diagram of a holographic storage device.

At this moment the storage density of demonstration systems is about half that of the best available disks (slightly over 100 MB/cm^2), while the read rate is in the order of 1 GB/s. The storage capacity might seem somewhat disappointing, but this is merely due to the mechanical steering mechanism used in beam steering and the refractive index of the medium used. Both factors of the storage device are crucial for the angular resolution and therefore for the density that can be obtained. Moreover, multiplexing data using different wavelengths from a tunable laser is still in its infancy. On maturity this will multiply the storage density by as many wavelengths as can be applied. It very much will depend on the development of other media, like rewritable DVD-type disks, whether holographic storage will become economically viable in the next 5 years.
What holographic storage makes particularly attractive in the area of data mining is the page-oriented way of storing data. One can correlate the data on pages at very high speed by overlaying stored data with a reference page. Only the differences between these pages show up and constitute a correlation measure for the reference and the data page.

References

- Amza, C., A.L. Cox, S. Dwarkadas, P. Keleher, H. Lu, R. Rajamony, W. Yu, W. Zwaenepoel. TreadMarks: Shared Memory Computing on Networks of Workstations. To appear in IEEE Computer (also: www.cs.rice.edu/~willy/TreadMarks/papers.htm)
- Baker, M. (ed.). (2000). Cluster Computing White Paper. www.netlib.org/benchmark/top500 www.dcs.port.ac.uk/mab/tfcc/WhitePaper/
- Barkes, J., M.R. Barrios, F. Cougard, P.G. Crumley, D, Marin, H, Reddy, T. Thitayanun. (1998). GPFS: A Parallel File System, IBM International Support Organization. www.redbooks.ibm.com
- Chandra, R., L. Dagum, D. Kohr, D. Maydan, J. McDonald, R. Menon. (2001). Parallel Programming in OpenMP. Morgan Kaufmann
- Cheung, D. (et al.). (1996). A Fast Distributed Algorithm for Mining Association Rules. 4th International Conference Parallel and Distributed Information Systems. IEEE Comp. Soc. Press, Los Alamitos, California, USA. pp31-42
- Flynn, M.J. (1972). Some Computer Organisations and their Effectiveness. IEEE Trans. on Computers **C-21** (9):948-960
- Geist, A., A. Beguelin, J. Dongarra, R. Manchek, W. Jaing, V. Sunderam. (1994). PVM: A Users' Guide and Tutorial for Networked Parallel Computing. MIT Press, Boston
- Gropp, W., S. Huss-Ledermann, A. Lumsdaine, E. Lusk, B. Nitzberg, W. Saphir, M. Snir. (1998). MPI: The Complete Reference **2**. The MPI Extensions, MIT Press, Boston

- Han, E.H., G. Kapyris, V. Kumar. (1997). Scalable Parallel Data Mining for Association Rules. Proceedings ACM Conf. Management of Data. ACM Press, New York, USA. pp277-288
- High Performance Fortran Forum. (1993). High Performance Fortran Language Specification. Scientific Programming **2** (13):1-170
- Hockney, R.W., C.R. Jesshope. (1987). Parallel Computers II. Adam Hilger, Bristol
- Horie, T., H. Ishihata, T. Shimizu, S. Kato, S. Inano, M. Ikesaka. (1991). AP1000 Architecture and Performance of LU Decomposition. Proceedings International Symposium on Supercomputing. Fukuoka. pp46-55
- James, D.V., A.T. Laundrie, S. Gjessing, G.S. Sohi. (1990). Scalable Coherent Interface. IEEE Computer **23** (6):74-77. http://sunrise.scu.edu
- OpenMP Forum. (1997). Fortran Language Specification **1.0**. http://www.openmp.org
- Snir, M., S. Otto, S. Huss-Lederman, D. Walker, J. Dongarra. (1998). MPI: The Complete Reference **1**. The MPI Core, MIT Press, Boston
- Sobol, I.M., (1994), A Primer for the Monte Carlo Method. Inst Mathematical Modeling, Moscow, Russia
- Spector, D.H.M. (2000). Building Unix Clusters. O'Reilly, Sebastopol, CA, USA
- Steen, A.J. van der. (1990). Exploring VLIW: Benchmark Tests on a Multiflow TRACE 14/300. Academic Computing Centre Utrecht. Technical Report TR-31
- Steen, A.J. van der. (1991). The Benchmark of the EuroBen Group. Parallel Computing **17**:1211-1221
- Steen, A.J. van der. (1993). Benchmark Results for the Hitachi S-3800. Supercomputer **10** (4/5):32-45
- Steen, A.J. van der. (ed.). (1995). Aspects of Computational Science. NCF, The Hague
- Steen, A.J. van der. (2000). An Evaluation of Some Beowulf Clusters. Technical Report WFI-00-07. Utrecht University, Department of Computational Physics. (Also available through www.euroben.nl, directory reports/)
- Sterling, T.L., J. Salmon, D.J. Becker, D.F. Savaresse. (1999). How to Build a Beowulf. The MIT Press, Boston
- www.hitachi.co.jp/Prod/comp/hpc/eng/sr1.html
- Zaki, M.J. (1999). Parallel and Distributed Association Mining: A Survey. IEEE Concurrency:14-25

6.4.2 Parallel data mining

Domenico Talia[1]

Motivations

Today the information overload is a problem like the shortage of information. Knowledge discovery in large data repositories can find what is interesting in them and represent it in an understandable way [Berry, 1997]. Mining large data sets requires large computational resources. In fact, data mining algorithms working on conventional computers on very large data sets take a very long time to get results. One approach to reducing response time is sampling. However, in some cases reducing data might result in inaccurate models, in some other cases it is not useful (e.g. outlier identification). The other approach is parallel computing. High performance computers and parallel data mining algorithms can offer the best way to mine very large data sets [Freitas, 1998; Skillicorn, 1999].

It is not uncommon to have sequential data mining applications that require several days or weeks to complete their task. Parallel computing systems can bring significant benefits in the implementation of data mining and knowledge discovery applications by means of the exploitation of inherent parallelism of data mining algorithms.

Data mining and parallel computing

The main goals of the use of parallel computing technologies in the data mining field are:

- performance improvements of existing techniques;
- implementation of new (parallel) techniques and algorithms;
- concurrent analysis with different data mining techniques and result integration to get a better model (that is more accurate).

There are three main strategies in the exploitation of parallelism in data mining algorithms:

- independent parallelism;
- task parallelism;
- SPMD[2] parallelism.

According to task parallelism (or control parallelism) each process executes different operations on (a different partition of) the data set. In SPMD parallelism a set of processes execute the same algorithm in parallel on different partitions of the data set and processes exchange partial results. Finally, independent parallelism is exploited; when processes are executed in parallel in an independent way, generally each process has access to the whole data set.

1 Dr D. Talia, talia@si.deis.unical.it, ISI-CNR, Institute of System Analysis and Information Technology, Italy, http://isi-cnr.deis.unical.it:1080/~talia

2 Single Program Multiple Data.

These three strategies are not necessarily exclusive for parallelizing data mining algorithms. They can be combined to improve both performance and accuracy of results. In combination with strategies for parallelization, different data partition strategies can be used:

- sequential partitioning: separate partitions are defined without overlapping among them;
- cover-based partitioning: some data can be replicated on different partitions;
- range-based query: partitions are defined on the basis of some queries that select data according to attribute values.

Parallelism in data mining techniques

In this paragraph for each data mining technique different parallelization strategies are outlined and some parallel data mining tools, algorithms or systems are cited.

Table 1
An overview of the main data mining tasks.

Data mining tasks	Data mining techniques
Classification	induction, neural networks, genetic algorithms
Association	Apriori, statistics, genetic algorithms
Clustering	neural networks, induction, statistics
Regression	induction, neural networks, statistics
Episode discovery	induction, neural networks, genetic algorithms
Summarization	induction, statistics

In Table 1 the main data mining tasks are listed and for each task the main techniques used to solve them are indicated. In the following sections we describe different approaches for parallel implementation of some techniques listed in Table 1.

Parallel decision trees (parallel induction)

Task parallel approach

According to the task parallelism approach one process is associated to each subtree of the decision tree that is built to represent a classification model. The search occurs in parallel in each subtree, thus the degree of parallelism P is equal to the number of active processes at a given time.
A possible implementation of this approach is based on farm parallelism in which there is a master process that controls the computation and a set of workers that are assigned to the subtrees.

SPMD approach

In the exploitation of SPDM parallelism each process classifies the items of a subset of data. The P processes search the whole tree in parallel using a parti-

tion of the data set *D/P*. The global result is obtained by exchanging partial results. The data set partitioning can usually be operated in two different ways:

– partitioning the tuples of the data set: (*D/P*) per processor;
– partitioning the *n* attributes of each tuple: *D* tuples of (*n/P*) attributes per processor.

In [Kufrin, 1997] a parallel implementation of the C4.5 algorithm that use the independent parallelism approach is discussed. Other significant examples of parallel algorithms that use decision trees are SPRINT discussed in [Shafer, 1996] and TDIDT (Top-Down Induction of Decision Trees) [Pearson, 1994].

Discovery of association rules in parallel

SPMD approach

In the SPMD strategy the data set is partitioned among the processors, but candidate itemsets are replicated on each processor. Each process *p* counts the partial support in parallel of the global itemsets on its local partition of the data set *D/p*. At the end of this phase global support is obtained by collecting all local supports.
The replication of the candidate itemsets minimizes communication, but does not use memory efficiently. Due to low communication overhead, scalability is good.

Task parallel approach

In this case both data set and candidate itemsets are partitioned on each processor. Each process *p* counts the global support of its candidate itemset on the entire data set *D*. After scanning its local data set partition, a process must scan all remote partitions for each iteration. The partitioning of data set and candidate itemsets minimizes the use of memory, but requires high communication overhead. Due to communication overhead this approach is not scalable, unlike the previous one.

Hybrid approaches

Combination of different parallelism approaches can be designed. For example, SPMD and task parallelism can be combined by defining clusters of processors composed of the same number of processing nodes. The data set is partitioned among the clusters, thus each cluster is responsible for computing the partial support of the candidate itemsets according to the SPMD approach. Each processor in a cluster uses the task parallel approach to compute the support of its disjoint set of candidates, scanning the dataset stored on the processors of its cluster. At the end of each iteration the cluster communicates to compute the global support.

The Apriori algorithm [Agrawal, 1994] is the best known algorithm for association rules discovery. Several parallel implementations have been proposed for this algorithm. In [Agrawal, 1996] two different parallel algorithms called Count Distribution (CD) and Data Distribution (DD) are presented. The first one is based on independent parallelism and the second one is based on task parallelism. In [Han, 1999] two different parallel approaches to Apriori called Intelligent Data Distribution (IDD) and Hybrid Distribution (HD) are presented. A complete review of parallel algorithms for association rules can be found in [Zaki, 1999].

Parallel neural networks

Neural Networks (NN) are a biology-inspired model of parallel computing. Supervised NN are used to implement classification algorithms and unsupervised NN are used to implement clustering algorithms. A lot of work on parallel implementation of neural networks has been done in the past. Theoretically, each neuron can be executed in parallel. But, in practice the grain of processors is generally larger than the grain of neurons, and the processor interconnection degree is restricted in comparison with neuron interconnection. Then a subset of neurons is generally mapped on each processor.

There are several different ways to exploit parallelism in a neural network:

- parallelism among training sessions: simultaneous execution of different training sessions;
- parallelism among training examples: each processor trains the same network on a subset of $1/P$ examples;
- layers parallelism: each layer of a neural network is mapped on a different processor;
- column parallelism: the neurons that belong to a column are executed on a different processor;
- weight parallelism: parallel execution of weight summation for connections of each neuron.

These parallel approaches can be combined to form different hybrid parallelization strategies. Different combinations can raise different issues to be faced for efficient implementation:

- interconnection topology;
- mapping strategies;
- load balancing among the processors;
- communication latency.

Typical parallelism approaches that are used for the implementation of neural networks on parallel architectures are:

– task parallelism;
– SPMD parallelism;
– farm parallelism.

Clementine is a parallel data mining system based on neural nets [McLaren, 1997]. In [Chattratichat, 1997] a task-parallel implementation of a back-propagation network is described and in [Rüger, 1997] a parallel implementation of a Self-organizing map is discussed. Finally, Neural Network Utility (NNU) [Bigus, 1996] is a neural network-based data mining environment that also has been implemented on a IBM SP2 parallel machine.

Parallel genetic algorithms

Genetic algorithms are used today for several data mining tasks such as classification, association rules, and episode discovery. Parallelism can be exploited in some phases of a genetic algorithm:
– population initialization;
– fitness computation;
– execution of the mutation operator,
without modifying the behavior of the algorithm in comparison to the sequential version.
The parallel execution of selection and crossover operations requires the definition of new strategies that modify the behavior (and results) of a genetic algorithm in comparison to the sequential version. The most frequently used approach is called global parallelization. It is based on the parallel execution of the fitness function and mutation operator. The other operation are executed sequentially. However, there are two possible SMPD variants:
– Each processor receives a subset of elements and evaluates their fitness using the entire data set.
– Each processor receives a subset of the data set and evaluates the fitness of every population element on its local subset.

Global parallelization can be effective when very large data sets are used. This approach is simple and has the same behavior of its sequential version, however, its implementations have not achieved very good performance and scalability, because of communication overhead.
Two different parallelization strategies that can change the behavior of the genetic algorithm are:
– *The island model (coarse grained)*: Each processor executes the genetic algorithm on a subset N/P of elements (subdemes) and periodically the best elements of a subpopulation are migrated towards the other the processors.
– *The diffusion model (fine grained)*: The population is divided in a large number of subpopulations composed of few individuals that evolve in parallel.

Several subsets are mapped on one processor. Typically elements are arranged in a regular topology (e.g. a grid). Each element evolves in parallel and executes the selection and crossover operations with the neighbor elements.

A very simple strategy is the independent parallel execution of *P* independent copies of a genetic algorithm on *P* processors. The final result is selected as the best one among the *P* results. Different parameters and initial populations should be used for each copy. In this approach there is no communication overhead. In this case the goal is not a higher performance, but a better accuracy. Some significant examples of data mining systems based on the parallel execution of genetic algorithms are *GA-MINER* [Flockart, 1996], *REGAL* [Neri, 1995], and *G-NET* [Anglano, 1997].

Parallel cluster analysis

Clustering algorithms arrange data items into several groups, called clusters so that similar items fall into the same group. This is done without any suggestion from an external supervisor, so classes are not given a priori, but must be discovered by the algorithm. When used to classify large data sets, clustering algorithms are very computing intensive.
Clustering algorithms can roughly be classified into two groups: hierarchical and partitioning models. Hierarchical methods generate a hierarchical decomposition of a set of *N* items represented by a dendogram. Each level of a dendogram identifies a possible set of clusters. Dendograms can be built starting from one cluster and iteratively this cluster is split until *N* clusters are obtained (divisive methods) or starting with the *N* clusters. At each step two clusters are merged until only one is left (agglomerative methods). Partitioning methods partition a set of objects into *K* clusters using a distance measure. Most of these approaches assume that the number *K* of groups has been given a priori.
Parallelism in clustering algorithms can be exploited both in the clustering strategy and in the computation of the similarity or distance among the data items, by computing on each processor the distance/similarity of a different partition of items.
Also in the parallel implementation of clustering algorithms the three main parallel strategies can be adopted.

Independent parallel approach

Each processor uses the whole dataset and it performs a different classification based on a different number of clusters. To get the load among the processors balanced a new classification is assigned to a processor that completed its task.

Task parallel approach

Each processor executes a different task that composes the clustering algorithm and cooperates with other processors exchanging partial results. For example, in partitioning methods processors can work on disjoint regions of the search space using the whole data set. In hierarchical methods a processor can be responsible of one or more clusters. It finds the nearest neighbor cluster by computing the distance among its cluster and the others. Then all the local shortest distances are exchanged to find the global shortest distance between two clusters that must be merged. The new cluster will be assigned to one of the two processors.

SPMD approach

Each processor executes the same algorithm on a different partition of the data set to compute partial clustering results. Local results are then exchanged among all the processors to get global values on every processor. The global values are used in all processors to start the next clustering step until a convergence is reached or a certain number of steps are executed. The SPMD strategy can be also used to implement clustering algorithms where each processor generates a local approximation of a model (classification) that at each iteration can be passed to the other processors that can use it to improve their clustering model.

In [Olson, 1995] a set of hierarchical clustering algorithms and an analysis of time complexity on different parallel architectures can be found. An example of parallel implementation of a clustering algorithm is P-CLUSTER [Judd, 1996]. Other parallel clustering algorithms are discussed in [Bruynooghe, 1989; Li, 1989; Foti, 2000]. In particular, in [Foti, 2000] an SPDM implementation of the AutoClass algorithm is described. The paper shows interesting performance results on distributed memory MIMD machines.

Architectural issues

In presenting the different strategies for the parallel implementation of data mining techniques we did not address architectural issues such as:

- distributed memory versus shared memory implementation;
- interconnection topology of processors;
- optimal communication strategies;
- load balancing of parallel data mining algorithms;
- memory usage and optimization;
- I/O impact on algorithm performance.

These issues (and others) must be taken into account in the process of parallelism exploitation in data mining algorithms. On the other hand, these issues offer several research opportunities for those interested in investigating paral-

lel implementation of data mining systems on different parallel architectures. The architectural issues are strongly related to the parallelization strategies and there is a mutual influence between the knowledge extraction strategy and the architectural features. For instance, an increase in the parallelism degree in some cases corresponds to an increase of the communication overhead among the processors. However, communication costs can be also balanced by the improved knowledge that a data mining algorithm can acquire from parallelization. At each iteration the processors share the approximated models produced by each one of them. Thus, each processor executes a next iteration using its own previous work and also the knowledge produced by the other processors. This approach can improve the rate at which a data mining algorithm finds a model for data (knowledge) and makes up for lost time in communication.

Research issues and directions

Parallel execution of different data mining algorithms and techniques can be integrated to obtain a better model not just to get high performance, but also high accuracy. Here we list some promising research directions in the parallel data mining area:

- Not just parallel algorithms, but environments and tools for interactive high performance data mining and knowledge discovery.
- Parallel text mining.
- Parallel and distributed web mining.
- Integration of parallel data mining with parallel data warehouses.
- Use of parallel computing in all the phases of the KDD process and support of efficient data warehouses.

Besides these very promising areas, we would like to mention the importance of the integrated use of clusters and grids for distributed and parallel knowledge discovery. Grid integrated clusters of computers that execute the same or different data mining or KDD algorithms can be seen as massive parallel computers that mine very large data sets. The development of software architectures, environments and tools for grid-based data mining will result in Grid-aware PDKD[1] systems that will support high performance data mining applications on geographically distributed data sources [Cannataro, 2000].

Conclusion

Applications in the area of data management and analysis show the highest growth rate among the applications developed on parallel computing machines [Strohmaier, 1997]. However, industrial and government data mining success stories tend not to be publicized. Parallel and distributed data mining will play a more and more important role for data analysis and knowledge extraction in several application contexts analysis of scientific data mining of commercial,

1 Parallel and Distributed Knowledge Discovery.

business and financial databases, data extraction and decision support for government and public departments. Data mining algorithms and underlying techniques can be parallelized to make them effective in the analysis of very large data sets. Several parallel strategies, algorithms, techniques, prototypes have been developed in the recent years. They allow researchers and end-users to mine large databases offering scalable performance. Nevertheless many promising research issues need to be faced and interesting directions must be explored. Data mining and in general knowledge discovery is an area in which parallel computing is used not only for quantitative computing, as occurs in many scientific computing applications, but also for qualitative computing.

References

- Agrawal, R., R. Srikant. (1994). Fast Algorithms for Mining Association Rules. Proceedings of the 20th International Conference on Very Large Databases, Santiago, Chile
- Agrawal, R., J.C. Shafer. (1996). Parallel Mining of Association Rules. IEEE Transactions on Knowledge and Data Engineering **8**
- Anglano, C., A. Giordana, G. Lo Bello, L. Saitta. (1997). A Network Genetic Algorithm for Concept Learning. Proceedings 7th International Conference Genetic Algorithms. pp434-441
- Berry, M., J.A.G. Linoff. (1997). Data Mining Techniques for Marketing, Sales, and Customer Support. Wiley Computer Publishing
- Bigus, J.P. (1996). Data Mining with Neural Networks. McGraw-Hill, New York
- Bruynooghe, M. (1989). Parallel Implementation of Fast Clustering Algorithms. Proceedings International Symposium On High Performance Computing. pp65-78
- Cannataro, M., D. Talia. (2000). Parallel and Distributed Knowledge Discovery on the Grid: A Reference Architecture. Proceedings of the 4th International Conference on Algorithms and Architectures for Parallel Computing (ICA3PP). Hong Kong, World Scientific Publ.
- Chattratichat, J. (et al.). (1997). Large Scale Data Mining: Challenges and Responses. Proceedings Third International Conference on KDD
- Flockart, I.W., N.J. Radcliffe. (1996). A Genetic Algorithm-Based Approach to Data Mining. Proceedings KDD-96 2nd International Conference On Knowledge Discovery and Data Mining. AAAI Press. pp299-302
- Foti, D., D. Lipari, C. Pizzuti, D. Talia. (2000). Scalable Parallel Clustering for Data Mining on Multicomputers. Proceedings of the 3rd International Workshop on High Performance Data Mining HPDMoo-IPDPS. LNCS, Springer Verlag, Cancun
- Freitas, A.A., S.H. Lavington. (1998). Mining Very Large Database with Parallel Processing. Kluwer Academic Publishers
- Han, E.-H., G. Karypis, V. Kumar. (1999). Scalable Parallel Data Mining for

Association Rules. IEEE Transactions on Knowledge and Data Engineering
- Judd, D., K. McKinley, A.K. Jain. (1996). Large-Scale Parallel Data Clustering. Proceedings International Conference On Pattern Recognition, Vienna
- Kufrin, R. (1997). Generating C4.5 Production Rules in Parallel. Proceedings 14th National Conference on Artificial Intelligence (AAAI-97). AAAI Press
- Li, X., Z. Fang. (1989). Parallel Clustering Algorithms. Parallel Computing **11**:275-290
- McLaren, I., E. Babb, J. Bocca. (1997). DAFS: Supporting the Knowledge Discovery Process. Proceedings 1st International Conference Practical Applications of Knowledge Discovery. The Practical Application Company. pp179-190
- Neri, F., A. Giordana. (1995). A Parallel Genetic Algorithm for Concept Learning. Proceedings 6th International Conference Genetic Algorithms. pp436-443
- Olson, C.F. (1995). Parallel Algorithms for Hierarchical Clustering. Parallel Computing **21**:313-1325
- Pearson, R.A. (1994). A Coarse-Grained Parallel Induction Heuristic. In: H. Kitano, V. Kumar, C.B. Suttner. (eds.). Parallel Processing for Artificial Intelligence **2**:207-226. Elsevier Science
- Rüger, S.M. (1997). Parallel Self-Organizing Maps. Proceedings PCW'97
- Shafer, J., R. Agrawal, M. Mehta. (1996). SPRINT: A Scalable Parallel Classifier for Data Mining. Proceedings 22nd International Conference on Very Large Databases (VLDB-96), Bombay
- Skillicorn, D. (1999). Strategies for Parallel Data Mining. IEEE Concurrency **7**
- Strohmaier, E., J.J. Dongarra, H.W. Meuer, H.D. Simon. (1997). Industrial Application Areas of High-Performance Computing. Proceedings of HPCN Europe '97. Springer Verlag. pp3-10. LNCS 1225
- Zaki, M.J. (1999). Parallel and Distributed Association Mining: A Survey. IEEE Concurrency **7**:14-25

6.4.3 Relational data mining

Marc de Haas[1], *Nico Brandt*[2]

Introduction

Current data mining approaches are enhanced through cutting edge research from statistics, machine learning, visualization, and database management. There are advances in mining of new data formats and structures, and successes with new and hybrid algorithms. New data mining user environments make data mining easier for domain experts to use and provide methods for automated preprocessing of data. Web mining brings new challenges for investigating web content, structure and usage. New research explores the possibilities of data mining in distributed environments. In other words, a lot is happening in the field of data mining.

In this chapter we will cover one of the new approaches in data mining, and give a view on trends and opportunities in this field.

This chapter is organized as follows. The next section describes data mining algorithms that can mine on complex structured objects. The construction of algorithms that can mine on structured objects is not beneficial unless end-users are able to configure and run these algorithms conveniently, understand the results, and use the results to solve their business problems. This topic is covered in the section on new user environments.

Practical experience has shown that rich sources of information are provided by an increasingly large number of distributed and heterogeneous databases. This presents an urgent need for effective and efficient mining techniques that can handle multiple databases. Advances in this field are presented in the section on distributed environments.

Business issues, like the impact of the Internet and E-commerce, which have resulted in the accumulation of vast amounts of transactional and demographic data are described in the section on business issues.

New data mining algorithms

The field of data mining is rich in new developments. There have been great achievements in the creation of fast algorithms. Traditional data mining algorithms, such as C4.5 [Quinlan, 1993] and others [Agrawal, 1994], can discover patterns that can be expressed in attribute-value languages which have the expressive power of propositional logic. In other words, these data mining algorithms can mine on data that is represented in one single table. These languages are limited and do not allow for representation of complex structured objects and relations among objects or their components. Real-world data is often too complex to represent in an attribute-value language. As an example we can take data that is stored in a relational database. In most relational data

1 Drs M. de Haas, Perot Systems Nederland B.V., Amersfoort, The Netherlands, http://www.perotsystems.nl

2 N. Brandt, Nico.Brandt@ps.net, Syllogic Innovations, Perot Systems Nederland B.V., Amersfoort, The Netherlands, http://www.perotsystems.nl

models structured objects exists, represented by tables with relations (foreign key constraints). Advances in the field of machine learning, such as multi-relational data mining [Knobbe, 1999] and Inductive Logic Programming (ILP) [Džeroski, 1996], have lead to techniques that can mine on structured objects. A natural extension of mining on tables with relations is mining on objects in an object oriented environment. In this section we focus on multi-relational data mining and object oriented data mining.

Multi-relational data mining

A multi-relational data mining algorithm finds patterns in a set of tables with relations between these tables. To illustrate the added value of this approach consider the following example.

We have a simple data model that consists of two tables, *Person* and *Speed Tickets*[3]. There is a foreign key constraint *Speed Tickets(PersonId)* that references *Person(PersonId)*. This means that there is a 1 to 0..n relation between *Person* and *Speed Tickets*.

Table 1
Example of a Person database.

Person				
PersonID	**DOB**	**Gender**	**Last Name**	**FirstName**
100001	06-30 1969	Male	Hare	Marc
123214	05-28 1967	Female	Brink	Mascha
231003	02-11 1971	Male	Firestone	Nico

Table 2
Example of a Speed Tickets database.

Speed Tickets					
TicketID	**PersonID**	**Time**	**Speed (km/h)**	**Allowed Speed**	**Speed Violation**
90123	123214	20:23 06-01 2001	67	50	17
90127	123214	10:22 06-04 2001	59	50	9
90135	231003	10:37 06-12 2001	150	120	30
90145	231003	18:33 07-05 2001	64	50	14
90158	231003	09:53 07-15 2001	55	50	5
90201	100001	11:19 07-15 2001	155	120	35
90311	100001	19:30 07-17 2001	153	120	33
90355	123214	10:15 07-24 2001	155	120	35
90366	231003	23:04 08-09 2001	162	120	42

A rule that a multi-relational data mining algorithm could learn from this data set is:

If a Person has two or more speed tickets for driving more than 30 km/h too fast

3 This is an example, we don't want to encourage any person to violate speed limits.

Then the gender is male.

When we use a traditional attribute-value data mining algorithm, we have to propositional this data model in the preprocessing phase, with as result one single table. Notice that simply joining the two tables will not produce a desirable result, because this changes (the distribution) of the analyzed population. We have to summarize the data in the *Speed Ticket* table by using aggregation functions like count, avg[1], sum, min, max, predominant, and join the result with the *Person* table. The resulting table has three rows and has, for example, the following attributes*: PersonId, DOB, Gender, Last Name, First Name, number of speed tickets, max speed violation, min speed violation, avg speed violation*. With this table we can learn rules like:

If a Person has more than three speed tickets or the avg speed violation is more than 20.3 km/h
Then the gender is male.

But an attribute-value data mining algorithm can not learn a rule like the one we learned in the example with the multi-relational data mining algorithm. It is not possible to find subgroups of the *form two tickets with speed violation of more than 30 km/h*.
From the example we can also see another advantage of a multi-relational data mining algorithm. The manual preprocessing step that transforms the available data in a propositional representation is not necessary. This does not mean that summarizing data is not allowed. If a summarized feature is important from the application domain point of view, this feature can be added in a multi-relational data mining case precisely the same as with the attribute-value data mining case. Practical experience has shown that the time spend on preprocessing can take from 50% up to 80% of the entire data mining process. When the data model contains a lot of tables and relations, considerable time can be saved in the preprocessing phase, when using a multi-relational data mining algorithm.

Object oriented data mining

The step from multi-relational data mining to object oriented data mining seems natural. Object oriented databases, like ObjectStore, Caché and Matisse, have become mature and are gaining popularity. They have the advantage that they can store the concepts, specified in an object oriented modeling language directly, and can query these concepts efficiently. When using an object oriented database (OODB), data is stored in an environment in which there is more information about the structure of the stored concepts. In other words, domain knowledge of the data is contained in the storage structure.
Let us consider the following simple example: we have three relational tables,

1 Average.

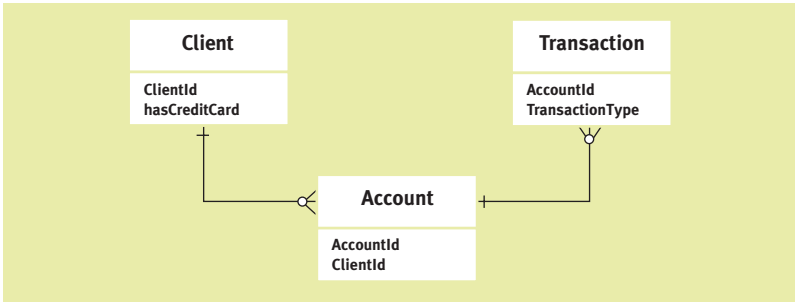


Figure 1
Data model from Client, Accounts and Transaction tables.

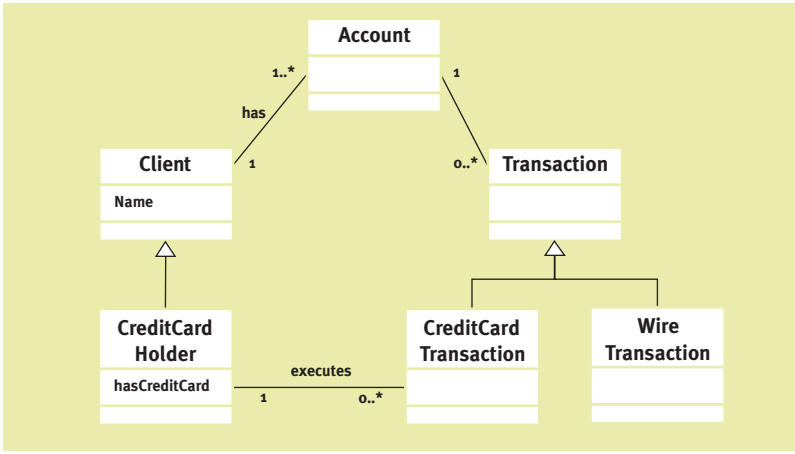


Figure 2
Object oriented form of the model from figure 1 with inheritance.

Client, *Account* and *Transaction*. One of the columns of the *Client* table is named *hasCreditCard* and stores information about the possession of a credit card of a client (yes or no). The client table has a relation with the *Account* table and the *Account* table has relations with the *Transaction* table.

The Transaction table has a column named *TransactionType* with possible nominal values *WT* which stands for wire transaction and *CT* which stands for credit card transaction. From this structure we cannot see that only credit card holders will execute credit card transactions. Of course this rule can be learned by a multi-relational data mining algorithm, but it is a trivial rule, which is well-known by the domain expert. In an object oriented database we could store the data using an inheritance structure. Known relations can be added more intuitively, which makes the learning process more efficient.

The computational price

Making the step from traditional data mining to multi-relational data mining is computationally expensive. Multi-relational and object oriented data mining algorithms search in a much larger hypothesis space, in order to find a suitable model for the given data set, than the traditional algorithms. Fortunately there are techniques that speed up the model building.

Consider, for example, the notion that a multi-relational data mining algorithm has to execute a large number of similar queries. Independent execution of these queries means a lot of redundant computations. These redundancies can be removed by integrating the similar queries in a so-called query-pack [Blockeel, 2000].

Taming the computational complexity of multi relational data mining algorithms based on inductive logic programming (ILP), for example, have been achieved by syntactically restricting the First Order predicate Logic (FOL) on which ILP is based. This technique, combined with search heuristics, has delivered some good results. Another, promising approach for applying ILP techniques is to use ILP in less expressive, substructural logics with nicer computational behavior than FOL [Adriaans, 2000a].

New user environments

Domain understanding and description of activities are essential phases in the data mining process. Data mining environments have to offer the user a graphical language that can be used to identify, organize and model the knowledge about the data and the data mining process and supports its reuse for similar situations.

Ideally, a complete data mining process model, like CRISP-DM [4], forms the basis of a data mining environment. A graphical representation of the data mining process and the application domain enables the domain expert and data mining expert to get a better overview and provides them a basis for communication. Extending the data mining environment by providing an automatic translation from the (graphical) specification of the domain to direct executable preprocessing actions on the data and suitable parameters for the data mining algorithm is an enhancement that really makes a difference. It enables the data mining expert to focus on the features and relations that she or he wants to use in the learning process without the burden of planning and implementation of low level actions on the data, such as windowing or summarizing the data using SQL or stored procedures.

Indications are that improving the interface to the data mining process can lead to considerable speed-up [MiningMart, 1999].

Using UML for specifying domain knowledge

The Object Oriented Modeling Language UML [5] has grown into the de facto visual language for specifying, constructing and documenting the artifacts of information systems. The object oriented features like inheritance, encapsulation, association and aggregation enable intuitive and compact specification of the data mining process and its entities. Vendors of data mining environments have noticed the popularity of UML as specification language and begin to offer UML like interfaces to enable users to model their data mining tasks.

4 CRISP-DM stands for Cross Industry Standard Process for Data Mining. For more information see http://www.crisp-dm.org

5 The Unified Modeling Language (UML) specification is maintained by the Object Management Group (OMG). (OMG) is an open membership, not-for-profit consortium that produces and maintains computer industry specifications for interoperable enterprise applications. Their best-known specifications include CORBA, OMG IDL, IIOP, UML, the MOF, CWM and Domain Facilities in industries, such as healthcare, manufacturing. Fore more information visit their web site at http://www.omg.org

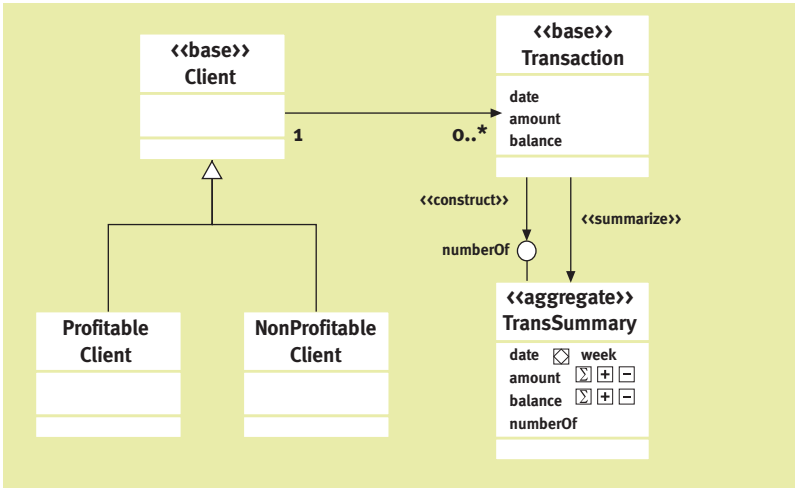


Figure 3
UML class diagram for data mining.

In the context of the data mining process the UML class diagrams can be used to give a static representation of the concepts a user wants to use in the learning process and the relation between these concepts. For modeling dynamic elements of the data mining process, like sequences of operations and iterations, UML sequence diagrams can be used. A good example is the MiningMart (see [MiningMart, 1999]) data mining environment which introduces extensions on UML that meet the specific needs for specification and visualization of the data mining process. In the example of the class diagram in Figure 3, we can see that in the application domain the distinction between *Profitable Client* and *Non Profitable Client* is important, and that the concept *Transactions summarized per week* is the interesting level of detail for the *Transaction* concept. The example of the sequence diagram in Figure 4 shows a high level description of the data mining process as it is defined in CRISP-DM.

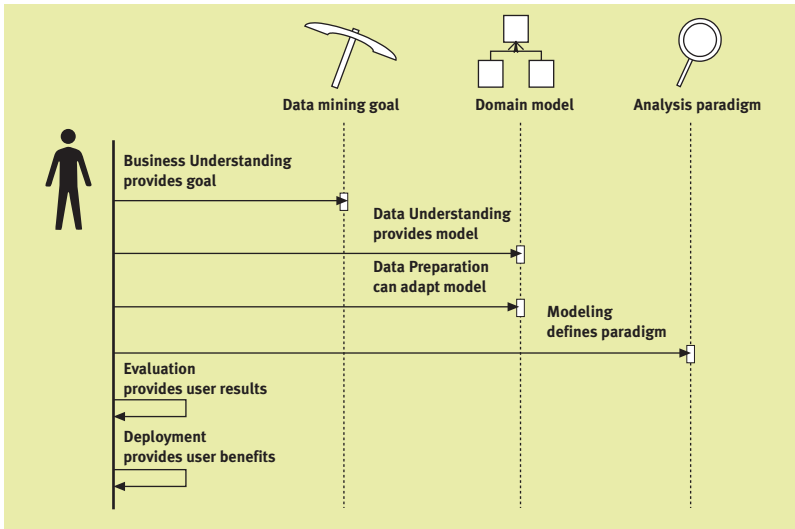


Figure 4
Data mining process according to the CRISP-DM model.

Distributed environments

A natural extension to the new techniques in data mining is mining in a distributed environment. Practical experience has shown that rich sources of information are provided by an increasingly large number of distributed and heterogeneous databases.

Distributed database environments can be modeled using a UML like specification language. This specification can be used directly for data mining, when an object oriented data mining algorithm is used. Concepts, such as *Consumer* and *Product Type* for example, can be specified on a global level. Mappings from these concepts can be specified to the different databases using dependencies, associations and inheritance. This task can be assisted by using techniques such as attribute equivalence theory [Dai, 1998], quantitative measure of relevance [Liu, 2000], text mining, ontology learning and grammar induction [Adriaans, 2000 b].

Business issues

Most industries including financial, manufacturing, telecommunications, energy and other utility companies have been experiencing enormous growth over the last decade. Growth has moved simultaneously in two directions:

- *Size*: companies have grown large as a result of mergers and acquisitions, new business models and new sales channels. This growth has resulted in more employees, more customers, more transactions, more distributed IT environments and more (internal) logistics.
- *Complexity*: forces behind increases in complexity are increasing quality standards, specialization within production processes, more sales channels and new technologies. Customers require products, services and business models fitting their needs.

This increase in size and complexity results in more and more structured and distributed data. Traditional data mining is capable of handling these large amounts of data, but handling the distributed environment without a lot of human effort is not possible.

The new data mining methods, multi-relational and object oriented, and new user environments explained in this article are new ways to ensure that organization can learn from all this information, by capturing the customer's perspective.

For instance, virtually all major organizations utilize basic usage analysis tools to understand their web site traffic, a few organizations couple this information with for instance their CRM environment, but even fewer extend those basic analysis capabilities with data mining methods to understand customer behavior and only a couple of organizations use the knowledge found to serve their customer better through individual adaptive web pages or mailings.

Another example is dot.com organizations. Being fast and smart are the keys to success in the dot.com arena, especially as competition heats up from bricks-and-mortar and clicks-and-mortar competitors. Dot.coms have the advantage of vast amounts of data without vast amounts of physical infrastructure. They collect information as a natural result of doing business electronically. This information must, however, be coupled with E-business intelligence solutions — whether internally or via extranets — otherwise these advantages in speed and intelligence will be lost. Never before has an industry had the quantity and quality of information that is available to dot.coms. It is only with the application of new data mining methods and user environments that this information can be used for competitive advantage. A key to dot.com success must be the appropriate use of E-business intelligence.

References

See also (included on the CD-rom): De Raedt, L., H. Blockeel, L. Dehaspe, W. van Laer. Three companions for data mining in first order logic. From: Relational Data Mining (S. Džeroski, and N. Lavrač, eds.), © Springer Verlag, 2001, pp. 105-139.

- Adriaans, P.W., E. de Haas. (2000a). Grammar Induction as Substructural Inductive Logic Programming
- Adriaans, P.W., M.H. Trautwein, M.R. Vervoort. (2000b). Towards High Speed Grammar Induction on Large Text Corpora. Proceedings of SOFSEM 2000. To Appear.
- Agrawal, R., R. Srikant. (1994). Fast Algorithms for Mining Association Rules. VLDB'94. pp487-499
- Blockeel, H., B. Demoem, L. Dehaspe, G. Janssens, J. Ramon, H. Vandecasteele. (2000.) Executing Query Packs in ILP. Proceedings of the 10th International Conference in Inductive Logic Programming. Volume **1866** of Lecture Notes in Artificial Intelligence:60-77. Springer Verlag
- Date, C.J., H. Darwen. (2000). Foundation for Future Database Systems. The Third Manifesto. Second Edition. Addison-Wesley
- Haas, E. de. (2001). Logics for Information Systems. ILLC Dissertation Series 2001-03
- Džeroski, S. (1996). Inductive Logic Programming and Knowledge Discovery in Databases. In: Tayyad, Piatetsky-Shapiro, Smyth, Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI Press/The MIT Press. pp117-152
- Huan, L., L. Hongjun, J. Yao. (2000). Towards Multidatabase Mining: Identifying Relevant Databases. IEEE Transactions on Knowledge and Data Engineering. IEEECS Log Number 105570
- Dai, H. (1998.). An Object Oriented Approach to Schema Integration and Data

Mining in Multiple Databases. Proceedings of the Technology of Object-Oriented Languages and Systems-Tools **24**
- MiningMart is an IST Research Program Funded by the European Committee. (1999). Proposal Number 11993, started 1999
- Morisio, M., G.H. Travassos, M.E. Stark. (2000). Extending UML to Support Domain Analysis. Proceedings of the Fifteenth IEEE International Conference on Automated Software Engineering
- Knobbe, A.J., H. Blockeel, A. Siebes, D.M.G. van der Wallen. (1999). Multi-Relational Data Mining. Proceedings of Benelearn '99
- Knobbe, A.J., A. Siebes, H. Blockeel, D. van der Wallen. (2000). Multi-Relational Data Mining, Using UML for ILP. Proceedings of PKDD 2000
- Lavrač, N., S. Džeroski. (1994) Inductive Logic Programming: Techniques and Applications. Ellis Horwood, New York (CD-rom) or http://www.ai.ijs.si/5asodzeroski/ILPBook/
- Quinlan, J.R. (1993). C4.5 Programs for Machine Learning. Morgan Kaufmann, San Mateo

6.4.4 Meta-learning

Christophe Giraud-Carrier[1], *Jörg Keller*[2]

Abstract

Data mining is the process of discovering useful knowledge in data. Central to this endeavor are data preprocessing and model selection, where users must choose: 1) transformations to apply to the raw data and 2) the learning algorithms best suited for the mining task at hand.

Since there is no 'cure-all' answer to these questions, business users must either proceed by trial and error or hire expertise. Neither solution is completely satisfactory. Instead, a number of researchers have suggested that learning itself can be applied to these issues, leading to a form of meta-learning. In this chapter, we introduce the concept of meta-learning, provide a brief historical overview and discuss current work in this area. An extensive list of references is included.

Introduction

Over the past decade, interest in Machine Learning (ML) and Data Mining (DM) technologies, particularly in the area of classification and prediction, has been growing rapidly in industry and commerce. Techniques have successfully started their transition from research laboratories to the real world and the number of fielded applications has increased steadily [Langley, 1995]. Many commercial data mining tools are already available, making the technology more readily usable. However, such tools remain of limited use to end-users who are not experts in machine learning/data mining. Most successful applications are custom-designed and the result of skilful use of human expertise or costly trial and error. This is due, in part, to the fact that ML/DM systems are non-trivial and their number keep increasing with no systematic method to discriminate among them. Hence, current data mining tools are only as powerful/useful as their users. They do provide multiple algorithms within a single, integrated system, but the selection of which algorithm is most suitable for a given application, as well as the possible combinations of these algorithms, are external to the system and left entirely up to the user's intelligence. This is clearly unsatisfactory for the non-expert end-users who wish to access a much-needed technology. Automatic and systematic guidance is required.

Automatic guidance in model selection, model combination and data transformation requires meta-knowledge. The cumulative expertise gained from ML/DM research and the conclusions of past comparative studies may serve as useful sources of prior meta-knowledge. Such prior knowledge can readily be augmented through the use of inductive (meta-)learning techniques. In particular, meta-learning offers an automatic way of inducing meta-knowledge from

1 C. Giraud-Carrier, cgc@elca.ch, ELCA Informatique SA, Lausanne 13, Switzerland

2 J. Keller, joerg.keller@daimlerchrysler.com, Daimler Chrysler AG, Data-Mining Solutions FT3/AD, D-89013 Ulm, Germany.

experience as well as revising prior meta-knowledge. With increasingly more models and techniques to choose from, meta-learning seems a prerequisite to the successful industrial/commercial take-up of ML/DM technology.
Meta-learning is machine learning at the meta-level. Whereas, at the base-level, the focus is on a specific learning task (e.g. credit rating, mine-rock discrimination, fraud detection), the focus at the meta-level is on the general learning problem. That is, meta-learning is about improving, by means of learning, the performance of the application of machine learning, i.e. learning to learn[3].

History and state of the art

Whilst it is impossible to give an exhaustive account of the work relevant to meta-learning, the following lists in roughly chronological order several of the most representative contributions.

- STABB [Utgoff, 1986] can be viewed as an early precursor of meta-learning systems, since it was the first to show that a learner's bias can be adjusted dynamically.
- VBMS [Rendell, 1989] is a relatively simple meta-learning system that learns to choose one of three symbolic learning algorithms as a function of only two dataset characteristics: the number of training instances and the number of features.
- Further work characterized and investigated the extensive role of data character as a determiner of system behaviour in empirical concept learning [Rendell, 1990]. The main contribution was the consideration of concepts as functions or surfaces over the instance space, which 1) leads to an useful characterization of complexity based on shape, size and concentration, and 2) allows systematic artificial data generation. The results, which focused on measures of complexity, showed that shape and especially concentration have significant effects. The results, however, are based on only two learning algorithms: ID3 and PLS1.
- The MLT project (ESPRIT Nr. 2154) focused on the practice of machine learning and produced a toolbox consisting of a number of learning algorithms, datasets, standards and know-how. Considerable insight was gained into many important learning issues. MLT produced a sophisticated tool, Consultant-2, capable of guiding the user in tool selection. Unfortunately, Consultant-2's knowledge base is static and contains information on only 10 learning algorithms, none of which are connectionist. Several improvements were suggested and reflected in the specification of Consultant-3, including guidance in data preprocessing [Sharma, 1993]. However, to the best of our knowledge, Consultant-3 was never implemented.
- The StatLog project (ESPRIT Nr. 5170) extended VBMS by considering a larger number of dataset characteristics, together with a broad class of candidate models and algorithms for selection [StatLog, 1991; Brazdil, 1994a;

3 We use the term here in a somewhat broader sense than its original definition by [Thrun, 1997].

Brazdil, 1994b; Gama, 1995]. The aim was to characterize the space in which a given algorithm achieves positive generalization performance. StatLog produced a thorough empirical analysis of machine learning algorithms and models. Some meta-learning for model selection was also attempted with bias restricted to accuracy only.

- JAM [Chan, 1996] takes a different approach to meta-learning. The system learns from partitioned data by using a set of base level classifiers and combining the learned results. The meta-learning task consists of learning from the predictions of a set of classifiers on common training data. Essentially, a meta-classifier is trained with the set of predictions of several base-level classifiers as input.
- An approach similar to that of JAM is found in [Ting, 1997]. The difference lies in the fact that the base-level classifiers use the same learning algorithm, but are trained on multiple, independent data batches. The focus is thus on data variation rather than on model variation.
- MetaL [Widmer, 1996a; Widmer, 1996b; Widmer, 1997] is also concerned with data variation, but as a time-dependent feature. The system performs on-line detection of concept drift with a single base-level classifier. The meta-learning task consists of identifying contextual clues, which are used to make the base-level classifier more selective with respect to training instances for prediction. Attributes and features that are characteristic of a specific context are identified and contextual features are used to focus on relevant examples, i.e. only those instances that match the context of the incoming training example are used as basis for prediction.
- MRL [Schmidhuber, 1996] represents a notable departure from the traditional use of classifiers at the object level. MRL is concerned with reinforcement learning, rather than supervised learning.
- Statistical modeling has also been applied to meta-learning. In particular, a statistical meta-model to predict the expected classification performance of 11 learning algorithms as a function of 13 data characteristics has been designed [Sohn, 1999]. The 13 data characteristics include basic descriptive statistics (e.g. number of classes) as well as multivariate statistics (e.g. mean skewness) and derived statistics (e.g. square root of the ratio of the number of feature variables to the number of training examples). 17 to 18 of the 22 datasets used in StatLog are used in this work to fit the statistical meta-model. A number of useful conclusions regarding the impact of certain characteristics on performance are drawn.
- The CRISP-DM project (ESPRIT Nr. 24.959) aimed to develop a cross-industry standard process for data mining [Chapman, 1999] and to embody this in a support tool, which provides user guidance based on the process model. CRISP-DM covers the entire data mining process from definition of business objectives and their translation into data mining goals through data access,

preparation, exploration and modeling, to results interpretation, assessment against objectives, deployment and documentation. Although not addressing meta-learning directly, this project is clearly relevant.

- The METAL project (ESPRIT Nr. 26.357), still running at the time of writing, focuses on 1) discovering new and relevant data/task characteristics, 2) using meta-learning to select either one in many or a ranking of classifiers, and 3) combining preprocessing with learning. The ultimate goal of the project is to embody meta-knowledge and an incremental meta-learner in a Web-enabled tool providing user guidance in algorithm selection. Details can be found in [METAL, 1998].
- Theoretical results, such as the NFL theorem and its consequences on meta-learning ([Schaffer, 1994; Wolpert, 1996a; Wolpert, 1996b; Wolpert, 1997; Wolpert, 2000]), help in identifying limitations and opportunities for meta-learning. For example, the latest analysis of [Wolpert, 2000] shows that for almost every single target, the generalization error of any two learning algorithms is almost exactly identical (for zero-one loss). This emphasizes just how important prior information about the target is and that unless the set of allowed targets is restricted to an extremely small set, any pair of algorithms considered will perform in essentially the same way. Such results are clearly relevant to work on model selection.
- Extensive empirical studies, such as [Aha, 1992; Holte, 1993; Lim, 2000; Vilalta, 2000] confirm the theory and provide much insight into learning both as sources of direct meta-knowledge and as input to meta-learning. For example, the results of [Holte, 1993] show that 1-rules (rules that classify an object on the basis of a single attribute) achieve surprisingly high accuracy on many datasets. In particular, they almost match C4.5 results. Additionally, The results of [Lim, 2000] show that most algorithms adapt to noise quite well and the mean error rates of many algorithms are so similar that their differences are statistically insignificant. This suggests that the differences are also probably insignificant in practical terms, that is, criteria other than predictive accuracy should be considered in selecting algorithms.

Finally, a most insightful account of the issues surrounding model class selection (that should be considered by anyone who is serious about meta-learning) can be found in [Someren, 2000]. A number of methods for model class selection are outlined and a taxonomy of meta-learning study types is proposed. The paper also addresses the issue of data transformation and its relation to model class selection.

In the following section we expand on the METAL project and its main contributions to meta-learning research and data mining applicability.

METAL: An European perspective on meta-learning

METAL is an ambitious R&D project broadly aimed at the development of methods and tools for providing support to users of machine learning (ML) and data mining (DM) technology. In particular, a web-enabled prototype assistant system that supports users with model selection and method combination and guides them through the space of experiments is implemented, as depicted below.

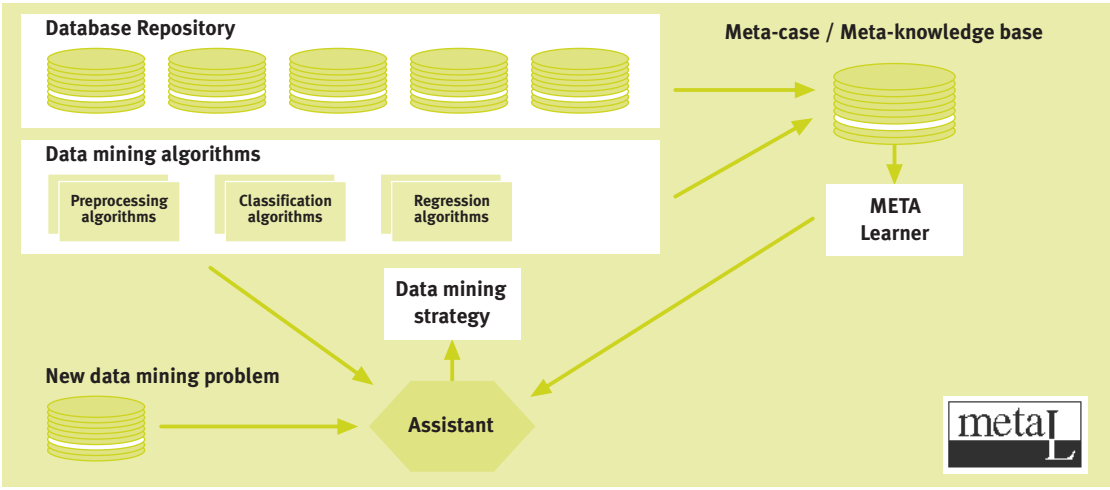


Figure 1
Metal meta-learning scheme.

The following briefly outlines some of the main contributions and design decisions of METAL. The expected effects, and the criteria by which the success of the project can be measured, are improved utility of data mining tools and in particular significant savings in experimentation time. A complete summary of the project's main findings will appear in [Brazdil, 2002].

Task characterization

The methods of task characterization developed and or studied within METAL rely on one or a combination of some form of the following three approaches:

- Statistical and information-theoretic characterization [Michie, 1994; Engels, 1998; Sohn, 1999; Köpf, 2000]. With this approach a number of statistics and information-theoretic measures are computed from the data set, e.g. number of attributes, skewness, class entropy, mutual information and signal-to-noise ratio. The basic assumption is that learning algorithms can be separated from one another along the dimensions offered by such information.
- Landmarking [Bensusan, 2000a; Pfahringer, 2000; Fürnkranz, 2000]. With this approach, the performances (i.e. cross validated predictive accuracy) of a number of simple and efficient learners on a given task are used to characterize it. The basic assumption is that the learners' space can be carved out into areas of expertise and that the performances of some (simple) learners

give an indication of the performances of other (more elaborate) learners.
– Model-(usually decision-tree)-based characterization [Bensusan, 1998; Bensusan, 2000b; Hilario, 2000a; Hilario, 2000b]. In this case a decision tree is induced from the data as well as a number of corresponding measures, such as depth, shape and balance. A task is then characterized by the induced decision tree (either directly or indirectly). The assumption of this approach is that decision-trees induced from data sets possess characteristics that are strongly dependent upon the data/task.

As with base-level learning, there are far more available (meta-)features than there are relevant ones. Hence, much effort needs also to be invested into the issue of feature selection [Kalousis, 2000; Todorovski, 2000b].

Model selection

METAL's work on model selection covers the more traditional 'choose 1 in *N*' paradigm [Bensusan, 2000a; Pfahringer, 2000], as well as a novel approach to combining several learning models by means of meta-decision trees [Todorovski, 1999; Todorovski, 2000a].
However, for its user-oriented meta-assistant, METAL has chosen a ranking strategy as the basis for its advice. The (meta-)learning task for ranking is different from the (meta-)learning task for finding a single best or for combining. In the former, the meta-model delivers not a single candidate, but an ordered list of algorithms, sorted from best to worst (based on some pre-defined criteria). One can argue that such rankings are more flexible and informative for users. Indeed, users are not restricted to a 'take it or leave it' kind of advice, with no clue whatsoever as to what to do, if the choice is to 'leave it' or if the 'take it' solution appears unsatisfactory. Rankings provide alternatives to users who may bring to bear on their final decision either their own expertise or any other criterion (e.g. financial constraints). Examples of meta-learning rankings are in [Nakhaeizadehl, 1997; Jammernegg, 1998; Berrer, 2000; Brazdil, 2000a; Brazdil, 2000b; Keller, 2000b; Soares, 2000].

Multi-criteria selection/ranking

Most work on model selection and or ranking has used predictive accuracy as the single selection criterion. Yet, there exist a number of other potentially useful (even important) selection criteria, such as computational complexity, comprehensibility and robustness [Müller, 1998; Giraud-Carrier, 1998].
To accommodate more than a single criterion, multi-criteria meta-learning approaches are emerging, such as [Nakhaeizadeh, 1997; Soares, 2000]. A rather promising approach is found in Data Envelopment Analysis (DEA) [Andersen, 1993; Paterson, 2000]. The method allows simultaneous scaling as well as translation invariance for inputs and outputs, and supports attributes with neg-

ative values directly. Hence, much preprocessing can be avoided. The following section describes some successful applications of DEA-based ranking.

Applications

In addition to a large number of controlled experiments on synthetic datasets and datasets from the UCI repository [Blake, 1998], METAL has been keen to extend its results to real business problems. Three of them are described here, all of them courtesy of DaimlerChrysler (DC).

Experimental design

The automotive industry is continually facing tougher competition on prices, product variety, quality and vehicle-development cycles. At the same time, tighter restrictions on fuel consumption and exhaust emissions are being imposed by governments worldwide. Meeting the requirements of the customers, whilst also complying with legal restrictions is a challenge that can only be accomplished thanks to high technology tools [Rezende, 1999].
Due to the high time pressure within the concept and development phases of a new product, the technical divisions demand quick prognosis methods and experimental design decision support. Based on DEA, a ranking tool named DCRanker was developed for the laboratory staff for ranking tests documented in DC data bases [Keller, 1999]. The multi-criteria ranking method developed has been a great help in the fast design of experiments to reduce time and costs.

Customer evaluation

Customer value is a common criterion for deciding on which prospects a company should spend its short acquisition budget. It is defined as an account's contributions to the well-being of a company. These can be either positive (benefit) or negative (costs). Researchers today agree that benefits and costs can be both monetary and non-monetary. However, the inclusion of non-monetary components is more often required than actually realized. Whenever researchers actually measure customer value, they mostly limit their research to monetary components. Since DEA can use differently scaled variables, it is useful to calculate a value score which covers both monetary and non-monetary components [Gelbrich, 2001].

Digital engineering

An implementation of meta-learning in digital engineering is intended at DC to produce a knowledge-based assistant for design engineers. There, a common information model (data, methods, tools, facilitation resources, know-how, etc.), which belongs to the product life-cycle, from the first product idea up to production, is available in an integrated form. An optimal interlocking of different assigned technologies is realized for the product and process development,

which leads to a clear decrease of the necessary development times and to an increase of the product and manufacturing process quality. Here, the issue for the fundamental knowledge-based engineering technology is to support the draft of vehicles, modules and aggregates with methods of knowledge management, knowledge processing and data mining, with the implementation of meta-learning, for an automatic selection of classification and regression methods [Keller, 2000a].

Summary and prospects

In today's fast-moving, data-rich business world, data mining has become a necessity to transform corporate data from a liability into an asset. Although simple in principle, the application of data mining is far from trivial in practice, due partly to the vast number of available techniques, each with its own specificity and applicability.

In this context, automatic decision-making methods which support practitioners in model selection and/or ranking and method combination need to be developed. Although human expertise is a great resource, it is often expensive, not always readily available, and subject to bias and personal preferences. As such expertise is generally acquired through experience, meta-learning is a promising complement to prior expert knowledge.

In this chapter, we have presented a brief history of meta-learning and highlighted the contributions of the most recent European project on this subject [METAL, 1998]. In particular, components and methods, developed in the context of a meta-learning assistant, have already shown a high application potential. A data characterization tool, called DCT, has been developed, together with two complementary ranking techniques: LIACC-Ranker and DCRanker. DCT and LIACC-Ranker have been implemented as SPSS Clementine's nodes, as well as within Weka (Weka is included on the CD-rom). For DCRanker, a stand-alone Windows-based version with a graphical user interface has been developed.

A beta version of a web dissemination site has been tested. On this site users may upload their datasets and receive a ranking of ten learning algorithms, based on either one of the ranking techniques described above, and using both predictive accuracy and training time as criteria. Furthermore, the user can run any of the ten algorithms on their dataset and obtain actual accuracy and time results. Once fully tested, this site will be available to the public from METAL's official web page [METAL, 1998].

Cleary, METAL's meta-learning advisor for model selection and combination is a very practical complement to the Cross Industry Standard Process in Data Mining (CRISP DM). Classification and regression learning tasks are rather common in daily business practice across a number of sectors. Therefore, the decision support offered by a meta-learning assistant is poised to be of great added value for data mining practitioners.

References

- Aha, D. (1992). Generalizing from Case Studies: A Case Study. Proceedings of the Ninth International Conference on Machine Learning (ICML-92)
- Andersen, P. N.C. Petersen. (1993). A Procedure for Ranking Efficient Units in Data Envelopment Analysis. Management Science **39** (10):1261-1264
- Bensusan, H. (1998). God Doesn't Always Shave with Occam's Razor – Learning When and How to Prune. Proceedings of the Tenth European Conference on Machine Learning (ECML-98)
- Bensusan, H., C. Giraud-Carrier. (2000a). Discovering Task Neighbourhoods Through Landmark Learning Performances. Proceedings of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD-2000)
- Bensusan, H., C. Giraud-Carrier, C. Kennedy. (2000b). A Higher-order Approach to Meta-learning. Proceedings of the ECML-2000 Workshop on Meta-learning: Building Automatic Advice Strategies for Model Selection and Method Combination
- Berrer, H., I. Paterson. J. Keller, J. (2000). Evaluation of Machine-learning Algorithm Ranking Advisors. Proceedings of the PKDD-2000 Workshop on Data-Mining, Decision Support, Meta-Learning and ILP. Forum for practical problem presentation and prospective solutions
- Blake, C.L, C.J. Merz. (1998). UCI Repository of Machine Learning Datasets. http://www.ics.uci.edu/~mlearn/MLRepository.html
- Brazdil, P., B. Henery. (1994a). Analysis of Results. In: D. Michie. (et al.). (eds.). Machine Learning, Neural and Statistical Classification. Chapter 10. Ellis Horwood.
- Brazdil, P., J. Gama, B. Henery. (1994b). Characterizing the Applicability of Classification Algorithms Using Meta-Level Learning. Proceedings of the Seventh European Conference on Machine Learning (ECML-94)
- Brazdil, P., C. Soares. (2000a). A Comparison of Ranking Methods for Classification Algorithm Selection. Proceedings of the Twelfth European Conference on Machine Learning (ECML-2000)
- Brazdil, P., C. Soares. (2000b). Ranking Classification Algorithms Based on Relevant Performance Information. Proceedings of the ECML-2000 Workshop on Meta-learning: Building Automatic Advice Strategies for Model Selection and Method Combination
- Brazdil, P., P. Flach, C. Giraud-Carrier. (eds.). (2002). Improving the Effectiveness of Data Mining via Meta-learning. In preparation
- Chan, P., S. Stolfo. (1996). On the Accuracy of Meta-Learning for Scalable Data Mining. Journal of Intelligent Information Systems **8**:3-28
- Chapman, P., J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth. (1999). CRISP-DM 1.0: Step-by-step data mining guide.

- Engels, R., C. Theusinger. (1998). Using a Data Metric for Offering Preprocessing Advice in Data-mining Applications. Proceedings of the Thirteenth European Conference on Artificial Intelligence (ECAI-98)
- Fürnkranz, J., J. Petrak. (2000). Two Remarks on Landmarking. Unpublished Manuscript
- Gama, J., P. Brazdil. (1995). Characterization of Classification Algorithms. Proceedings of the Seventh Portuguese Conference on Artificial Intelligence (EPIA-95)
- Gelbrich, K., I. Paterson, H. Berrer, J. Keller. (2001). Customer Evaluation as a Problem of Multi-criteria Ranking: An Application to the Financial Sector. Proceedings of the First SIAM International Conference on Data Mining
- Giraud-Carrier, C. (1998). Beyond Predictive Accuracy: What? ECML'98 Workshop Notes – Upgrading Learning to the Meta-Level. Model Selection and Data Transformation
- Hilario, M., A. Kalousis. (2000a). Building Algorithm Profiles for Prior Model Selection in Knowledge Discovery Systems. Engineering Intelligent Systems **8** (2)
- Hilario, M., A. Kalousis. (2000b). Quantifying the Resilience of Inductive Classification Algorithms. Proceedings of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD-2000)
- Holte, R.C. (1993). Very Simple Classification Rules Perform well on Most Commonly Used Datasets. Machine Learning **11**:63-91
- Jammernegg, W., M. Luptacik, G. Nakhaeizadeh, A. Schnabel. (1998). Ist ein fairer Vergleich von Data-Mining Algorithmen möglich? In: G. Nakhaeizadeh. (ed.). Data-Mining: Theoretische Aspekte und Anwendungen:225-240. Physica Verlag
- Kalousis, A., M. Hilario, M. (2000). Feature Selection for Meta-learning. In Proceedings of the Fifth Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD-2000)
- Keller, J., I. Holzer, S. Silvery. (1999). Using Data Envelopment Analysis and Cased-based Reasoning Techniques for Knowledge-based Engine-intake Port Design. Proceedings of the Twelfth International Conference on Engineering Design (ICED-99)
- Keller, J., V. Bauer, W. Kwedlo. (2000a). Application of Data Mining and Knowledge Discovery in Automotive Data Engineering. Proceedings of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD-2000)
- Keller, J., I. Paterson, H. Berrer. (2000b). An Integrated Concept for Multi-criteria Ranking of Data-mining Algorithms. Proceedings of the ECML-2000 Workshop on Meta-learning: Building Automatic Advice Strategies for Model Selection and Method Combination

- Köpf, C., C.C. Taylor, J. Keller. (2000). Meta-analysis: From Data Characterisation for Meta-learning to Meta-regression. Proceedings of the PKDD-2000 Workshop on Data-Mining, Decision Support, Meta-Learning and ILP. Forum for practical problem presentation and prospective solutions
- Langley, P., H.A. Simon. (1995). Applications of Machine Learning and Rule Induction. Communications of the ACM **38** (11):55-64
- Lim, T-S., W-Y. Loh, Y-S. Shih. (2000). A Comparison of Prediction Accuracy, Complexity and Training Time of Thirty-Three Old and New Classification Algorithms. Machine Learning **40**:203-228
- METAL (1998). A Meta-Learning Assistant for Providing User Support in Machine Learning and Data Mining. http://www.metal-kdd.org
- Michie, D., D.J. Spiegelhalter, C.C. Taylor. (eds.). (1994). Machine Learning, Neural and Statistical Classification. Ellis Horwood
- Müller, M., C. Hausdorf, J. Schneeberger. (1998). Zur Interessantheit bei der Entdeckung von Wissen in Datenbanken. In: G. Nakhaeizadeh. (ed.). Data-Mining: Theoretische Aspekte und Anwendungen:249-264. Physica Verlag
- Nakhaeizadeh, G., A. Schnabel. (1997). Development of Multi-criteria Metrics for Evaluation of Data-mining Algorithms. Proceedings of the Third International Conference on Knowledge Discovery and Data-Mining (KDD-97)
- Paterson, I. (2000). New Models for Data Envelopment Analysis, Measuring Efficiency with the VRS Frontier. Economics Series No. 84. Institute for Advanced Studies, Vienna
- Pfahringer, B., H. Bensusan, C. Giraud-Carrier. (2000). Meta-learning by Landmarking Various Learning Algorithms. Proceedings of the Seventeenth International Conference on Machine Learning (ICML-2000)
- Rendell, L., R. Seshu, D. Tcheng. (1989). Layered Concept-Learning and Dynamical Variable Bias Management. Proceedings of the Eleventh International Joint Conference on Artificial Intelligence (IJCAI-89)
- Rendell, L., H. Cho. (1990). Empirical Learning as a Function of Concept Character. Machine Learning **5**:267-298
- Rezende, F., G. Oliveira, R.C.G. Pereira, U. Hermsen, J. Keller. (1999). A Unified Database Interface for Multiple Heterogeneous Databases. Proceedings of the Conference on Engineering Federated Information Systems (EFIS-99)
- Schaffer, C. (1994). A Conservation Law for Generalization Performance. Proceedings of the Eleventh International Conference on Machine Learning (ICML-94)
- Schmidhuber, J., J. Zhao, M. Wiering. (1996). Simple Principles of Metalearning. Technical Report IDSIA-69-96, IDSIA, Switzerland
- Sharma, S., D. Sleeman, N. Graner, M. Rissakis. (1993). Specification of Consultant-3. Deliverable 5.7 of ESPRIT Project MLT (Nr. 2154). Ref: MLT/WP5/Abdn/D5.7

- Soares, C., P. Brazdil. (2000). Zoomed Ranking: Selection of Classification Algorithms Based on Relevant Performance Information. Proceedings of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD-2000)
- Sohn, S.Y. (1999). Meta Analysis of Classification Algorithms for Pattern Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, **21** (11):1137-1144
- Someren, M. van. (2000). Model Class Selection and Construction: Beyond the Procustean Approach to Machine Learning Applications. In: G. Paliouras, V. Karkaletsis, C.D. Spyropoulos. (eds.). Machine Learning and Applications, Springer Verlag
- StatLog. (1991). Comparative Testing and Evaluation of Statistical and Logical Learning Algorithms for Large-Scale Applications in Classification, Prediction and Control.
  http://www.newcastle.research.ec.org/esp-syn/text/5170.html
- Ting, K.M., B.T. Low. (1997). Model Combination in the Multiple-Data-Batches Scenario. Proceedings of the Ninth European Conference on Machine Learning (ECML-97)
- Todorovski, L., S. Džeroski. (1999). Experiments in Meta-level Learning with ILP. Proceedings of the Third European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD-99)
- Todorovski, L., S. Džeroski. (2000a). Combining Multiple Models with Meta Decision Trees. Proceedings of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD-2000)
- Todorovski, L., P. Brazdil, C. Soares. (2000b). Report on Experiments with Feature Selection in Meta-level Learning. Proceedings of the PKDD-2000 Workshop on Data-Mining, Decision Support, Meta-Learning and ILP. Forum for practical problem presentation and prospective solutions
- Utgoff, P.E. (1986). Machine Learning of Inductive Bias. Kluwer
- Vilalta, R., D. Oblinger. (2000). A Quantification of Distance-Bias Between Evaluation Metrics in Classification. Proceedings of the Seventeenth International Conference on Machine Learning (ICML-2000)
- Widmer, G. (1996a). On-line Meta-learning in Changing Contexts. MetaL(B) and MetaL(IB). Proceedings of the Third International Workshop on Multistrategy Learning (MSL-96)
- Widmer, G. (1996b). Recognition and Exploitation of Contextual Clues via Incremental Meta-Learning. Proceedings of the Thirteenth International Conference on Machine Learning (ICML-96)
- Widmer, G. (1997). Tracking Context Changes through Meta-Learning. Machine Learning, **27** (3):259-286
- Wolpert, D.H. (1996a). The Existence of A Priori Distinctions between Learning Algorithms. Neural Computation **7**

– Wolpert, D.H. (1996b). The Lack of A Priori Distinctions between Learning Algorithms. Neural Computation **7**
– Wolpert, D.H., W.G. Macready. (1997). No Free Lunch Theorems for Search. IEEE Transactions on Evolutionary Computation **1**
– Wolpert, D.H. (2000). Any Two Learning Algorithms Are (Almost) Exactly Identical. Proceedings of the ICML-2000 Workshop on What Works Well Where

6.4.5 Monitoring the results of the KDD process: an overview of pattern evolution

Steffan Baron[1], *Myra Spiliopoulou*[2]

Introduction

In the past ten years the number of algorithms and techniques appropriate for data mining tasks has grown tremendously. No matter which mining paradigm is considered there are many efficient approaches to get useful and valuable information from masses of data. Until recently, those techniques have been considered only for static datasets. However, the derived models and rules which reflect the logical dependencies in the dataset should be as dynamic as the underlying dataset itself. Rules discovered from a database are affected as records are added, modified or removed. Depending on the changes of the statistical properties of the dataset, existing rules may become invalid and new rules may emerge. Thus, we encounter a strong need to update observed rules as the dataset itself is updated.

In recent years, studies have emerged that address the issue of refreshing discovered knowledge. Most studies deal with association rules, although a small number of studies on sequence mining and cluster detection have also appeared. These studies concentrate on methods that actualize the mining results without reprocessing the whole dataset. A parallel stream of research investigates how changes in the dataset affect the validity of a rule or pattern, while still other studies try to detect when rules become invalid and a new mining session is due.

This study places those independent streams of research into a common context, that of monitoring the evolution of patterns in the presence of data change. In the next section, we concretize the concept of pattern evolution in the context of data mining, and propose a unifying framework for the studies that deal with this issue. In the subsequent sections, we discuss individual studies, organized across the dimensions of our framework. The last section concludes the study with an agenda for future research in this domain.

KDD patterns and their evolution

As new data is inserted into a dataset and existing data is modified or deleted, patterns that have been derived from the original dataset may become invalid, while new patterns may emerge. Since decision support is actually a continuous activity, monitoring of the changes in the data and of their effects on the patterns is indispensable, to ensure that reliable patterns are fed into the decision-making process.

In this section, we first describe the data mining results according to a very abstract schema. We use this schema to specify 'pattern evolution' in such a

1 Dr S. Baron, sbaron@wiwi.hu-berlin.de, Humboldt University Berlin, Institute of Information Systems, Berlin, Germany

2 Prof Dr Myra Spiliopoulou, myra@ebusiness.hhl.de, Department of E-Business, Handelshochschule Leipzig (HHL), Germany

generic way that both changes in the statistics of an association rule and in the contents of a cluster are reflected[3]. We then introduce the term 'monitor' to characterize tools that observe pattern evolution and detect changes.

The content and statistics of a pattern

The output of the knowledge discovery process is a set of patterns that reflect the application's mini-world, as described in the dataset. In general, a pattern consists of two parts: (i) its 'content' and (ii) its 'statistics' with respect to the dataset.

In the paradigm of association rules' discovery, a pattern is a set of items (content) that appear frequently in the transactions of the dataset [Agrawal, 1993]. It is accompanied by at least one statistical value, namely the support as ratio of the transactions containing the itemset to the total number of transactions. For patterns expressed as rules, additional statistics can be derived, such as the confidence, improvement and X^2-test value of the consequent with respect to the antecedent.

Frequent sequences discovered by a sequence miner can be similarly expressed as <content,statistics> pairs, whereby the pattern's content is an ordered sequence of items [Agrawal, 1994]. In later studies, the notions of frequent sequence and of large itemsets have been expanded in several ways. For example, several miners allow the members of an itemset or sequence to be sets of items themselves, so that individual items can be replaced by abstract concepts that represent whole ad hoc or predefined item groups [Agrawal, 1995; Mannila, 1996]. Bayardo concentrates on itemsets of maximal length [Bayardo, 1998], while Büchner et al. allow for the discovery of maximal length sequences during web usage mining [Baumgarten, 2000]. The Web Utilization Miner WUM discovers patterns that are sequences of trees, with statistics assigned at each node [Spiliopoulou, 1999], but there are still statistic values pertaining to the whole pattern. Thus, although the pattern's content can be a quite complex structure, frequent itemsets and sequences can be approximated by the <content,statistics> signature.

Clustering algorithms organize the records of a dataset into groups. Depending on the underlying clustering paradigm, group membership is described either intensionally, e.g. by the coordinates of a cluster's centroïd and the cluster's radius, or extensionally by giving the probability with which each record belongs to each cluster; e.g. the latter is the method of choice in AutoClass [Cheeseman, 1995]. This intensional or extensional description corresponds to the 'content'-part of our <content,statistics> pair for a cluster. The 'statistics'-part refers then to the cluster as a whole, and may capture such elements as the distribution of values for some property inside the cluster and in the whole population.

3 An earlier, more detailed version of this model can be found in [Spiliopoulou, 2000].

Data evolution and pattern evolution

We define 'data evolution' as: any modification in the dataset from which patterns have been derived. In a data warehouse, these modifications are restricted to new insertions. However, updates can also be performed, e.g. to correct errors, although they are rare. If data mining is being performed on an operational database or dataset, then we can expect that the data undergo deletions, updates and insertions. When the modified data become a substantial part of the new dataset, the patterns derived from the original one can no longer be assumed valid.

In the presence of data evolution, two courses of action are possible: a new data mining session can be launched over the whole dataset, or, an attempt can be made to identify the changes on the original patterns, namely the 'pattern evolution', preferably without processing the whole dataset. In fact, these two approaches are complementary: Extensive changes in the data should be expected to invalidate old patterns and make a new analysis necessary. On the other hand, the changes in the existing patterns are themselves important. For example, a company that introduces a new product onto the market is interested in the impact of this product on the sales of other, existing products; this cannot be done by merging old and new data, but only by analyzing them separately and comparing old and new patterns.

In the previous subsection, we described a pattern as a <content,statistics> pair. Data evolution can cause changes in the statistics of a pattern, e.g. a decrease in the support of a large itemset or a change in the population ratio of a cluster. Data evolution can also cause changes in the contents of a pattern. For example, records may migrate from one cluster to another, some of the items appearing in a large itemset may cease to be frequent. When the contents of a pattern change, an existential question emerges, namely to what extent a pattern is the same when its contents change. While one can claim that a large itemset or a frequent sequence is no longer the same if some of its members stop being frequent, a cluster does not 'die' when some of its members migrate. As we will see, techniques that monitor the evolution of association rules or frequent sequences can concentrate on verifying whether an existing rule disappears because parts of its content are no longer frequent and whether new rules emerge. On the other hand, techniques that monitor cluster change are forced to deal with changes in cluster contents that gradually result in new clusters replacing (parts of) old ones. In this study, we consider a collection of patterns derived from a dataset during a mining session. We use the term 'pattern evolution' for the following changes in the content of this collection:

a changes in the statistics of a pattern;
b discovery of a new pattern;
c changes in a pattern's content;
d disappearance of a pattern.

For association rules and frequent sequences, cases (c) and (d) are identical, while for clustering they should be treated separately.

Monitoring pattern evolution

In this study, we discuss methods that investigate the impact of data changes on the patterns derived from this data. We term these methods 'pattern monitors' or simply 'monitors', to stress their common property of monitoring the contents and the statistics of the patterns and registering the effects of data evolution on them.

Most methods that can be characterized as monitors according to our definition come from the emergent domain of incremental mining. Their goal is to update the mining results after a change in the dataset without processing the whole dataset. Depending on the type of mining results they monitor, they are confronted with one or more of the types of pattern evolution mentioned above.

A different group of methods concentrates on the first type of pattern evolution by monitoring the statistics of a pattern or a group of patterns. This group of monitors investigates the temporal aspects of pattern evolution by identifying time intervals, in which a pattern is valid, or by detecting interesting slopes in the curve of the pattern's statistics.

Related surveys

In [Roddick, 2000] there was a first attempt to place these methods in a common framework. However, the subject of [Roddick, 2000] was temporal mining. While data and pattern evolution is closely related to temporal mining, not all methods that conform to our definition of a 'monitor' can be claimed to perform temporal mining. Indeed, some of monitors do not perform data mining themselves, and are simply cooperating with a mining tool.

[Spiliopoulou, 2000] propose a framework for rule maintenance, in which rules are observed as temporal objects. Our generic model of a pattern as a <content,statistics> pair builds upon this earlier work. However, the focus of [Spiliopoulou, 2000] was on modeling evolving clusters, sequences and association rules in a unifying way, so that time series analysis and temporal mining could be applied on them. Here, our goal is rather to provide a unifying framework for the dispersed literature on the treatment of evolving mining results, thereby shedding light in the several aspects of pattern evolution. In this context, our study is unique and we hope that its integrative character will encourage information dissemination in this emergent research domain.

A unifying framework for pattern evolution monitors

Research contributions on the subject of pattern monitoring are dispersed across several domains and appear mostly as subordinate to some other research issue. For example, the study of [Chakrabarti, 1998] on the evolution of

correlated patterns assigns itself to the subject of discovering surprising patterns.

We introduce here a unifying framework for the contributions discussed in the following sections. In this framework, we observe pattern monitoring across five dimensions, described below and summarized in Table 1.

Mining paradigm

This dimension refers to the type of mining results being monitored. We have identified pattern monitors for the results of association rules discovery (Section 6.2.12), sequence mining (Section 6.2.10), cluster detection (Section 6.2.6) and inductive logic programming (ILP)[4] (Section 6.2.13).

Pattern treatment

Across this dimension we identify two types of monitors. The first type encompasses algorithms that refresh the patterns, thus turning a pattern into a series of time stamped <content,statistics> pairs. The second type is comprised of algorithms that concentrate on the changes themselves: they verify whether the patterns must be refreshed (using an algorithm of the first type). Some of these algorithms model a pattern as a temporal object and search for its maximal validity range.

Intervention of the monitor on the mining process

In this dimension, we consider the impact of the monitor on the conventional mining process. Conventionally, the analyst designs a mining session by specifying thresholds for some statistics like confidence and support; the miner discovers the rules that conform to these specifications. When the mining session is designed and driven by the monitor, the analyst's specifications must be combined with the requirements of the monitor. For example, a monitor may decide to decrease the specified thresholds and obtain also weak patterns, i.e. patterns that do not satisfy the given threshold yet, but may emerge after a data update. We term such patterns as internal ones, as opposed to external patterns such as those presented to the analyst. The extreme would be to generate all patterns from the dataset by switching off all statistical constraints.

According to this observation, we distinguish between:

– Monitors that do not intervene with the specifications that drive the mining process, i.e. they only use the external patterns normally produced by the miner, then
  $external\_patterns = internal\_patterns \subset all\_patterns$
– Monitors that require the discovery of additional patterns, i.e.
  $external\_patterns \subset internal\_patterns \subset all\_patterns$
– Monitors that generate all patterns, i.e.
  $external\_patterns \subset internal\_patterns = all\_patterns$

4 ILP is not a mining paradigm, but it induces patterns nonetheless.

Types of pattern evolution

This dimension reflects the four types of pattern evolution identified in the subsection on data evolution and pattern evolution. Some monitors investigate whether new rules emerge or old rules disappear from the rule base, as the data change. Other monitors consider rather an invariant rule base and study only changes of rule statistics.

Temporal semantics of pattern evolution

Change can be observed either as an event that occurs at a time point or as the border of the validity interval of the rule. Accordingly, a monitor may observe a rule as an entity defined across a series of time points or as an entity associated with a valid time interval.

Table 1 summarizes the framework: for each dimension, we specify the domain of values.

Table 1 [5]
The dimensions of the unifying framework.

Dimension	Value domain
Paradigm	association rules
	frequent sequences
	clusters
	generic (paradigm-independent)
Pattern treatment	type I: incremental miners
	type II: change detectors
Intervention of the monitor on the miner	external rules only
	internal rules
	all rules
Type of pattern evolution	change the pattern's statistics
	new pattern
	change the pattern's content
	disappearance of a pattern
Temporal semantics of pattern evolution	change conceived as event at a time point
	change conceived as a boundary of a validity interval

Usually, the increment database δ (delta) does not only consist of inserted items or transactions but also of deletions. However, without a loss of generality, an update on an existing record in *DB* is represented as deletion followed by an insertion.

5 Notation: since the used notation varies from study to study, we briefly introduce a set of variables, that we use in the rest of the survey. Table 2 gives an overview of used symbols, along with a short description for each of them.

Table 2
An overview of symbols used, along with a short description for each of them.

Symbol	Description
DB	original database
δ	increment database
$DB + \delta$	updated database
k	number of items
C_k	k - candidate itemset
L_k^{DB}, L_k^{δ}	large k -itemset in DB and in δ, respectively
$NB(L_k)$	negative border with respect to L_k, i.e. $C_k - L_k$

The last four lines of the table present symbols known from the field of association rules discovery, whereas k refers to the number of considered items in the set and the number of current iterations, C_k is the candidate itemset in the kth iteration, L_k^{DB} and L_k^{δ} are the large itemsets in DB and in δ in the kth iteration, and $NB(L_k)$ is the negative border with respect to the large k-itemset, i.e. a k-candidate itemset that did not have enough support to become large.

A guide through the rest of the paper: we use the first two dimensions of our framework to organize the rest of the paper. In the next section, we present contributions classified according to the second dimension of our framework as incremental miners. The section after that contains studies that detect data changes, and occasionally use an incremental miner to adjust the affect patterns accordingly. Within each section, we group the monitors by the mining paradigm they are designed for. We draw conclusions in the last section.

Incremental miners

We begin our survey on pattern monitoring by discussing incremental miners. An incremental miner is a miner that uses as input a set of patterns discovered during a mining session on a database DB and identifies emerging and obsolete patterns by incrementally processing an update of this database δ. Hence, all incremental miners are monitors that operate on externally scheduled mining sessions. All contributions we present observe pattern evolution as a series of changes at distinct time points, the time points of the mining sessions. Both changes in statistics and in content are recorded.

We organize this section according to the first dimension of our framework, i.e. the mining paradigm considered by each incremental-miner/monitor. The effects of the monitor on the conventional mining process (3rd dimension of our framework) are discussed separately.

Association rules discovery

The earliest incremental mining algorithms were designed for association rules. This mining paradigm is still the focus of the majority of contributions in this domain.

The family of FUP algorithms

We begin with three studies by [Cheung, 1996a; Cheung, 1996b, Cheung, 1997] for the incremental actualization of association rules.

They propose an algorithm FUP (Fast UPdate) designed to update association rules when a δ of new transactions is added to the original database *DB* (cf. Table 2). Basically, the framework of FUP algorithm is similar to the well-known algorithms Apriori [Agrawal et al., 1994] and DHP [Park, 1995]. In a number of iterations all the large itemsets are found. However, FUP uses information from previous mining sessions, including the old large itemsets L_k^{DB} and their support counts, to identify existing rules that become extinct and new rules that emerge due to δ. Thus, FUP supports changes in both the statistics and the contents of association rules, observing change as an event occurring at the time point of the mining session.

With respect to the third dimension of our framework, FUP is actually replacing the conventional miner but still uses the same statistical constraints that guided the analysis of *DB* and generates no additional intermediate results.

A major objective of FUP is the efficient updating of the rules. Performance improvement is achieved by pruning away many candidate rules by checking their supports in δ, which is in general much smaller than the original database. Moreover, only δ is scanned to filter out itemsets that are no longer large in the updated database [Cheung, 1996a].

Based on FUP, [Cheung, 1996b] propose two further algorithms, FUP* and MLUp (Multi-Level association rules Update) where the first one is an improved version of the original algorithm FUP, and the second algorithm is an adaptation of FUP* for the update of discovered multi-level association rules in relation databases. Finally, in [Cheung, 1997], they propose FUP_2, a generalization of FUP, which performs incremental mining in the presence of insertions into and deletions from the original database.

UWEP algorithm

[Ayan, 1999] propose a new algorithm UWEP (Update With Early Pruning), which employs a dynamic look ahead strategy to update the large itemsets in the updated database.

With respect to the dimensions of our framework, UWEP is very similar to the FUP algorithms: it updates discovered association rules, when a non-trivial number of new transactions δ is added to the database. It replaces the mining strategy, but otherwise generates only the patterns requested by the analyst

instead of relying on intermediate results. Pattern evolution encompasses changes in both statistics and content, including new and extinct rules. Also, similarly to the FUP group of algorithms, the maxim of UWEP is efficiency in performing the update.

The main performance advantage of UWEP is achieved by recognizing itemsets that will not remain large and pruning them away from the database *DB* and from and the increment δ at an early phase. In particular, UWEP prunes the supersets of a large itemset in *DB* as soon as it is known to be small in the updated database $DB + \delta$, without waiting until the *k*th iteration. This methodology yields a much smaller candidate set especially when the set of new transactions does not contain some of the old large itemsets.

A further performance gain is achieved by the processing mechanism: UWEP creates lists of transaction identifiers from the existing database *DB* and the update δ which then are used to calculate the large itemsets in the updated database $DB + \delta$, without re-scanning *DB*. Thus, UWEP scans the existing database at most once, and the new database exactly once, as opposed to the FUP algorithms that scan the database in each iteration.

An incremental variant of the partition algorithm

[Savasere, 1995] proposed the algorithm Partition for association rules discovery. In [Omiecinski, 1998] the authors introduce a variation of this algorithm, which refreshes association rules after an update to the original dataset.

If the dataset increment δ consists solely of insertions, a slight variation of the original Partition algorithm is used: the original database *DB* is treated as the one partition, and the increment δ as another partition. In the first phase, δ is scanned to identify all potentially large itemsets along with their support counts and the total number of transactions; the results of the previous mining session are re-used, so *DB* is not scanned again. In the second phase, Partition checks whether the large itemsets local to their partitions become large globally, i.e. in the updated database $DB + \delta$. This requires the scanning of *DB* once and of the increment δ once more. If a candidate set is large in both partitions it is large globally. Since the partitions are non-overlapping, the global support count is simply the sum of its local equivalents.

If δ contains also deletion operations, the algorithm must scan *DB* in the first phase, too, but can still use the old patterns to prune itemsets that do not remain large.

The incremental variant of Partition shares the same framework characteristics with UWEP and the FUP family, as well as the goal of updating the patterns efficiently, whereby 'efficiency' translates in a minimized number of scans over *DB* and δ.

Incremental mining using a negative border

[Thomas, 1997] propose an algorithm for knowledge maintenance, based on the notion of negative borders, as introduced by [Toivonen, 1996]. Informally, the negative border of the large k-itemsets $NB(L_k)$ consists of those candidates generated from the large (k -1)-itemsets, which did not have enough support counts to be classified as large, i.e. C_k - L_k.

The algorithm of Thomas et al. distinguishes between insert and delete operations on the database increment δ. First, it computes the large itemsets in δ, using a conventional association rules' discovery algorithm like Apriori [Agrawal, 1994] or Partition [Savasere, 1995]. Simultaneously, the algorithm counts the support for the itemsets L_k^{DB} and $NB(L_k^{DB})$ in δ. If an itemset is not large in $DB + \delta$ it is removed from the large itemsets, whereby the negative border is recomputed. Each large itemset in δ, is checked whether it gets enough support to move it from the negative border to the new large itemsets in the updated database. Then, if the union of the large itemsets and the negative border in $DB + \delta$ is not equal to the union of the large itemsets and the negative border in DB, the entire database is scanned once to compute the new negative border.

The algorithm of [Thomas, 1997] and the similar approach developed independently by [Feldman, 1997] at about the same time, are incremental miners that incorporate a conventional mining algorithm, which is customized for incremental processing of the updated dataset. They observe pattern evolution as change in pattern statistics of content, including new and extinct patterns. Change is an event occurring at the time point of the mining session, i.e. at the time point of analyzing DB and δ together.

Dissimilarly to the incremental miners described thus far, the monitors of [Thomas, 1997; Feldman, 1997] are based on an internally maintained set of patterns, which is a clear subset of the mining results requested by the analyst. This set of internal patterns (cf. 3rd dimension of our framework in the section on a unifying framework for pattern evolution monitors) is comprised by the mining results and the negative border. While other incremental miners reduce performance overheads by minimizing the number of scans over the data, the monitors of [Thomas, 1997; Feldman, 1997] achieve performance gains. However, it should be noted that this time gain is traded against the space overhead of retaining the border and the time overhead of updating it.

Sequence analysis

All previous algorithms were devoted to incremental association rules' discovery. Wang et al. propose the incremental discovery of sequences from evolving data [Wang, 1996; Wang, 1997]. Their algorithm is a sequence miner that automatically refreshes the patterns, whenever the dataset is updated.

At the core of the method of Wang et al. we find a method for the efficient maintenance of sequential data. They observe a database as a long sequence and

use a variation of a suffix tree called 'dynamic suffix tree' to index it. Then, the problem of discovering sequential patterns can be decomposed into finding the frequent substrings with respect to a given minimum support threshold, and gene rating frequent patterns from these substrings with respect to a given minimum confidence threshold. To solve this problem, it is sufficient to index all the substrings together with their support counts and position information; then, frequent sequences can be read off the tree representation immediately.
Data updates require the modification of the suffix tree. A suffix is only affected by an update if it contains a part of the updated substring and is not properly contained in a leaf. Such suffixes are called splitters. Based on this, they propose an algorithm which consists of two steps. In the first step, all splitters affected by a delete operation are removed, and references and dangling links are adjusted. In the second step, all splitters affected by an insert operation are included, and the support counts of nodes affected by deletions and insertions are updated. Unlike the position, insertion and deletion affects only the address of characters in the disk pages containing inserted or deleted characters.
The evolution of frequent sequences is monitored by (a) updating the suffix tree on the fly as the dataset changes and (b) extracting the frequent sequences from it. In terms of our framework in the section on a unifying framework for pattern evolution monitors, this algorithm is an incremental miner that is activated whenever the data change. Obviously, this concerns only step (a), while step (b) can be initiated by the user. Pattern evolution is then recorded as a series of the time points at which step (b) is performed. At each such time point, emerging new frequent sequences are recorded while sequences that are no longer frequent disappear. For all remaining sequences, the statistics are refreshed as the result of step (a).

Clustering algorithms

We close our discussion on incremental miners with a study on the incremental update of previously detected clusters by [Ester, 1998]. They propose an incremental clustering algorithm for a data warehouse: IncrementalDBSCAN is based on DBSCAN, a density-based clustering algorithm for metric databases, i.e. databases with a distance function for pairs of objects. The goal of IncrementalDBSCAN is the identification of the impact of data updates on the cluster contents and the adjustment of the clusters.
The key idea of density-based clustering is that for each object of a cluster its neighborhood (defined by a given radius) must contain a user-specified minimum number of objects. An object, for which this condition holds, is called the core object; a cluster is determined by any of its core objects. Border objects are located along a cluster's border, while noise objects do not belong to any cluster. Changes in the cluster members may thus shift a core object to another cluster, change a cluster's border or reduce the size of an object's neighborhood

below the minimum. IncrementalDBSCAN uses a set of heuristics that determine the impact of each update operation on each cluster.
Due to the density-based nature of DBSCAN, the insertion or deletion of an object affects the clustering (scheme) only in the neighborhood of this object. IncrementalDBSCAN examines which part of an existing clustering is affected by an update of the database and adjusts the clusters accordingly, whereby clusters may extinct or be merged and new clusters can appear.
IncrementalDBSCAN is not performing clustering in the conventional sense. However, the authors prove formally that its output is the same as for the DBSCAN clustering algorithm. Hence, IncrementalDBSCAN can be conceived as an incremental miner, which updates a clustering (scheme) by adjusting the clusters, after an increment δ over the original database *DB*. In its role as monitor, IncrementalDBSCAN perceives pattern evolution as content change only, thereby distinguishing among new clusters, extinct clusters and changes in cluster contents. Similarly to the previously described incremental miners, change is conceived as an event occurring at the time point of the new mining session.

Change detectors

The monitors in the section on incremental miners are characterized by their emphasis on updating the patterns efficiently and by the integration of the monitoring mechanism with the mining mechanism that discovers the new patterns, taking old ones into account.
In this section, we discuss monitors that concentrate on detecting change. The monitors of this type form a less homogenous group than incremental miners, showing variations across different dimensions of our unifying framework. We present them according to the same dimension as for section 3, namely the mining paradigm.

Association rules discovery

Incremental miners for association rules focus on updating the rules after a data change. Change detectors concentrate on identifying the modifications that are expected to affect the rules.

The DELI change detector

Lee et al. apply sampling techniques to detect changes that may affect previously discovered association rules [Lee, 1997, Lee, 1998]. They argue that all proposed methods for maintaining discovered association rules must examine not only the changed part but also the unchanged part in the original database. Since the unchanged part may be very large, the test can be time-consuming. Even if the increment δ is quite large, its impact on the rule set may be negligible, thus yielding the analysis redundant.
The monitor DELI (Difference Estimation for Large Itemsets) avoids this problem

by sampling. Sampling is used to estimate the difference between the association rules in a database before and after the database is updated. It is based on a distance measure which computes the symmetric difference between the old large itemsets and the new large itemsets, whereby the new large itemsets are discovered from the drawn sample. The estimated difference determines whether there should be an update on the mined association rules or not. If the estimated difference is smaller than a user-specified threshold, then the rules in the original database are still a good approximation to those in the updated database. The analyst can accumulate more updates before actually updating the rules, thereby avoiding the overheads of updating the rules too frequently.
In the context of our framework, DELI is not a miner. In fact, it uses the miner FUP2 to modify the patterns if this turns to be necessary. Thereby, no intervention with the miner takes place. With respect to the types of pattern evolution supported, DELI concentrates on changes in the contents of the patterns; effects on the statistics are handled by FUP_2. These changes are perceived as events on the time axis, corresponding to the time point(s) of the sampling process.

Monitoring a time series of association rules' statistics

A very particular type of monitor has been proposed in [Chakrabarti, 1998]. He introduced a new notion of surprising temporal patterns in the area of association rules discovery, and algorithms to find such patterns. In this context, a pattern is a change in the statistics of an association rule over time. They argue that once the analyst is already familiar with prevalent patterns in the data, the greatest incremental benefit is likely to be from changes in the relationship between item frequencies over time, whereas the degree of surprisingness is determined by the strength of the change in the statistics. They develop a notion of interest based on the number of bits needed to encode a itemset sequence using a specific coding scheme that they design. In this scheme it takes relatively few bits to encode sequences of itemsets that have a steady correlation between items. Conversely, a sequence with large code length hints at a possibly surprising correlation. The surprise value of the itemset is related to the difference or ratio between both code lengths. Moreover, their analysis produces, in a formal information-theoretic sense, the best segmentation of time for the interesting itemsets, based on how the relationship between items is changing.
In our terminology, the algorithm of [Chakrabarti, 1998] monitors changes in the statistics of association rules as the underlying data change. These rules can be discovered with any appropriate miner; the monitor does not intervene during their generation.
Dissimilarly to incremental miners, the time points of change are not given. Whilst DELI applies sampling to detect data updates that affect rule contents

and statistics, the monitor of [Chakrabarti, 1998] partitions the time axis into such intervals, that the rule statistics change dramatically between two consecutive intervals. With respect to the fourth dimension of our framework, the only type of pattern evolution observed concerns the pattern's statistics. With respect to the temporal semantics of pattern change, the partitioning of the time axis can be observed both as the generation of time points that indicate change and as the identification of intervals, among which changes in the rule's statistics occur.

Validity time intervals for association rules

The following study done by [Chen, 1999] aims at the discovery of valid time intervals of association rules. It focuses on two mining problems for temporal features of given, i.e. previously discovered, association rules: (a) finding all interesting contiguous time intervals during which a specific association holds, and (b) finding all interesting periodicities that a specific association has.
Given a time-stamped database of transactions and a known association rule they try to find all possible totally ordered time intervals during which the given association holds. These intervals have a given granularity and are not decomposable. They propose an algorithm called LISeeker which searches for all the longest intervals, i.e. the largest possible set of contiguous time intervals of a given association rule in a database over a time domain, during which this association satisfies minimum support and confidence. They also investigate the set of regular intervals in cycles, during each of which a given association holds, and propose an algorithm PIDeriver which can discover them.
In terms of our framework, they focus on valid time intervals and periodicities in the context of association rules, i.e. when changes to the data invalidate existing association rules. Actually, rather than dealing with incremental updation of associations, they consider the question when a given rule holds. The algorithms apply on existing rules which are used to determine their validity ranges. The proposed technique monitors changes of the statistics, no additional rules are discovered.

Logic programming and decision planning

In the following section, we introduce a study done by Pechoucek in the area of inductive logic programming and decision planning (DP) [Pechoucek, 1999]. They investigate the issue of gradual evolution of configuration meta-knowledge within a knowledge-based system. They distinguish between strong updates, i.e. re-computation of the entire inference knowledge base, and weak updates, i.e. re-computation of only the relevant parts of the inference knowledge base, and experiment with their frequencies to figure out what is the efficiency of the system after a sequence of weak updates. They apply two machine learning methodologies, inductive logic programming and explanation based

generalization within the framework of decision planning, on the problem of knowledge maintenance, and compare them by means of a case study for a configuration problem. While ILP is used to describe the training example, to express the induced knowledge, and to formulate background knowledge, DP is a declarative knowledge representation methodology based on proof planning. Both methods are compared to determine which knowledge is easier to maintain.

For ILP, they proved the weak update to be simple to achieve due to the used knowledge representation. Adding a new positive example takes just one single step, and the system exhibits an acceptable, linear decrease of efficiency with respect to the number of consecutive weak updates. They have found a simple method to estimate the frequency of strong updates in dependence on the requested efficiency of the target system. On the contrary, for DP, the time needed for the weak update is considerably higher than in the case of ILP. Moreover, as DP requires strategic inference knowledge acquired from the user in order to formalize an initial decision graph, its utilization is envisaged mainly in areas where human expertise is available.

As mentioned before, ILP is actually not a mining paradigm. However, since it induces patterns, and deals with the incremental update of derived knowledge it is relevant in this context. Regarding the fifth dimension of our framework, changes of the meta-knowledge are propagated against the knowledge base at specific time points.

Generic techniques

In the following subsection we introduce two studies done by [Ganti, 1999; Ganti, 2000] which are applicable on a variety of data mining techniques.

In their first study the difference between two datasets is investigated in order to quantify the difference between 'interesting' characteristics of two datasets [Ganti, 1999]. The proposed FOCUS framework can be used to compute an interpretable, qualifiable deviation measure between two datasets, whereas the difference is expressed in terms of the model the datasets induce. Their central idea is that a broad class of models can be described in terms of a structural component and a measure component, where the structural component identifies the interesting regions of the model and the measure component summarizes the subset of the data that is mapped to each region. They show how FOCUS can be used to identify subsets of data that lead to interesting changes in the discovered model. Moreover, the framework allows comparison of specific parts of two models and defines a set of operators to discover regions where the difference between two datasets is 'interesting'.

Their framework covers a wide variety of models including frequent itemsets, decision tree classifiers, and clusters. Patterns derived from DB and δ are used to determine the difference between the two datasets, whereas δ needs not

necessarily to be an update. In this regard, FOCUS is more generic and can be applied on any two datasets. Changes in the contents of a rule are observed as events at time points, and the causing data changes are identified.
Rather than the rule' s contents, [Ganti, 2000] investigate temporal aspects on the data level. They propose a general framework DEMON (Data Evolution and MONitoring) to deal with the problem space of systematic data evolution, and provide algorithms for the maintenance of derived models.
They introduce a new dimension for discovering patterns, called data span dimension, which allows user-defined selections of a temporal subset of the database. The model maintenance algorithms they propose take this new dimension into account. Interestingly, they distinguish between data that evolve systematically, for example in a data warehouse, and data that evolve non-systematically, as in a transactional database. Based on this, they allow selection constraints on the data span dimension which selects a set of data blocks the mining algorithm is applied on, whereas different blocks may span different time units. This way, the analyst is able to mine only that part of the data he is interested in, or that is relevant to a given task.
Though introduced on the specific example of maintenance of frequent itemsets and clusters, their generic algorithm can be instantiated for any suitable data mining technique. Regarding our third dimension no intermediate results are produced, because user-given thresholds for minimum support and confidence are used. On the contrary, additional constraints can be specified by the analyst to limit the data mining process to data from a specific time interval. Then, the discovered rules satisfy the statistical constraints for that particular time interval.

Conclusions

Huge amounts of data from different application domains are available, and the development of efficient data mining algorithms has become a major research challenge. However, since the data is rather dynamic than static, there is also a strong need to update rules and patterns that have been discovered from the data. In the last few years there have emerged a number of studies that deal with the problem of rule validity when data is added or deleted, mainly in the area of association rules discovery, and the goal of this study was to give a comprehensive overview of these works, positioning them in a unified framework.
Summarizing the introduced contributions, on the one hand, we have incremental miners, i.e. algorithms that are able to update previously discovered rules as data change, most of them building on their static counterparts. These works apply mainly in the area of incremental update of previously discovered association rules and aim at the efficiency of the proposed technique. On the other hand, we have monitors that investigate when patterns or rules are affected by updates to the dataset. These algorithms often use existing mining algorithms,

or they simply monitor the data rather than mining it. Mostly, these works represent frameworks to measure changes in data characteristics.
In the investigation of temporal features, two approaches have emerged: the first method uses user-given constraints on the time domain to discover patterns belonging to the specified time interval, and the second method uses given patterns to determine time intervals in which this rule holds. Alternatively, a combined approach could be employed to discover patterns, along with their valid time intervals from scratch. There are a few emerging studies in this area, but this might be a direction for future work.
Most of the studies considered, use given partitions *DB* and δ as starting point, and user-given thresholds for minimum support and confidence when updating patterns. Moreover, they concentrate on the contents of rules at given time points, i.e. whether new rules emerge or old rules become extinct when data is updated.
But there are also studies focusing on the statistical properties of rules. They draw time series of statistical parameters, and monitor their course when the underlying dataset changes. These works have rather invariant contents, and observe the evolution of a rule as change of its statistics. However, still missing is a framework that is capable to deal with all four types of change a rule may undergo, as mentioned above. This might also be a direction for future work.

Acknowledgement

Thanks to Gerrit Riessen for comments on the article.

References

– Agrawal, R., T. Imielinski, A.N. Swami. (1993). Mining Association Rules between Sets of Items in Large Databases. In: P. Buneman, S. Jajodia (eds.). Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data. pp207-216. ACM Press
– Agrawal, R., R. Srikant. (1994). Fast Algorithms for Mining Association Rules in Large Databases. In: J.B. Bocca, M. Jarke, C. Zaniolo. (eds.). VLDB'94, Proceedings of the 20th International Conference on Very Large Data Bases. pp487-499. Morgan Kaufmann
– Agrawal, R., R. Srikant. (1995). Mining Sequential Patterns. Proceedings of the International Conference on Data Engineering. Tapei, Taiwan
– Ayan, N.F., A.U. Tansel, E. Arkun. (1999). An Efficient Algorithm To Update Large Itemsets With Early Pruning. Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp287-291. ACM
– Baumgarten, M., A.G. Buchner, S.S. Anand, M.D. Mulvenna, J.G. Hughes. (2000). Navigation Pattern Discovery From Internet Data. WEBKDD'99: Workshop on Web Usage Analysis and User Profiling. pp70-87. Springer Verlag

- Bayardo, R.J. Jr. (1998). Efficiently Mining Long Patterns from Databases. SIGMOD'98. pp85-93. Seattle, WA
- Chakrabarti, S., S. Sarawagi, B. Dom. (1998). Mining Surprising Patterns Using Temporal Description Length. In: A. Gupta, O. Shmueli, J. Widom. (eds.). VLDB'98. pp606-617. Morgan Kaufmann
- Cheeseman, P., J. Stutz. (1995). Bayesian Classification (AutoClass): Theory and Results. Knowledge Discovery in Data Bases II. AAAI Press/The MIT Press
- Chen, X, I. Petrounias. (1999). Mining Temporal Features in Association Rules. Proceedings of the 3rd European Conference on Principles of Data Mining and Knowledge Discovery. Lecture Notes in Computer Science:295-300. Springer Verlag
- Cheung, D.W., J. Han, V.T. Ng, C.Y. Wong. (1996a). Maintenance of Discovered Association Rules in Large Databases: An Incremental Updating Technique. ICDE'96
- Cheung, D.W., V.T. Ng, B.W. Tam. (1996b). Maintenance of Discovered Knowledge: A Case in Multi-Level Association Rules. KDD'96
- Cheung, D.W., S.D. Lee, B. Kao. (1997). A General Incremental Technique for Maintaining Discovered Association Rules. DASFAA'97, Melbourne, Australia
- Ester, M., H-P. Kriegel, J. Sander, M. Wimmer, X. Xu. (1998). Incremental Clustering for Mining in a Data Warehousing Environment. Proceedings of the 24th International Conference on Very Large Data Bases. pp323-333. Morgan Kaufmann
- Feldman, R., Y. Aumann, A. Amir, H. Mannila. (1997). Efficient Algorithms for Discovering Frequent Sets in Incremental Databases. Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'97), Tucson, Arizona, USA
- Ganti, V., J. Gehrke, R. Ramakrishnan. (1999). A Framework for Measuring Changes in Data Characteristics. Proceedings of the Eighteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. pp126-137. ACM Press
- Ganti, V., J. Gehrke, R. Ramakrishnan. (2000). DEMON: Mining and Monitoring Evolving Data. Proceedings of the 15th International Conference on Data Engineering. pp439-448. IEEE Computer Society
- Lee, S.D., D.W. Cheung. (1997). Maintenance of Discovered Association Rules: When to Update? ACM-SIGMOD Workshop on Data Mining and Knowledge Discovery (DMKD-97), Tucson, Arizona
- Lee, S.D., D.W. Cheung, B. Kao. (1998). Is Sampling Useful in Data Mining? A Case in the Maintenance of Discovered Association Rules. Data Mining and Knowledge Discovery **2** (3):233-262
- Mannila, H., H. Toivonen. (1996). Discovering Generalized Episodes using Minimal Occurences. Proceedings of the 2nd International Conference

KDD'96. pp146-151
- Omiecinski, E., A. Savasere. (1998). Efficient Mining of Association Rules in Large Databases. Proceedings of the British National Conference on Databases. pp49-63
- Park, J.S., M.S. Chen, P.S. Yu. (1995). An Effective Hash Based Algorithm for Mining Association Rules. In: M.J. Carey, D.A. Schneider. (eds.). Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data. pp175-186. ACM Press
- Pechoucek, M., O. Stepankova, P. Miksovsky. (1999). Maintenance of Discovered Knowledge. Proceedings of the 3rd European Conference on Principles of Data Mining and Knowledge Discovery. Lecture Notes in Computer Science:476-483. Springer Verlag
- Roddick, J.F., M. Spiliopoulou. (2000). A Survey of Temporal Knowledge Discovery Paradigms and Methods. Accepted in IEEE Trans. of Knowledge and Data Engineering. To Appear
- Savasere, A., E. Omiecinski, S. Navathe. (1995). An Efficient Algorithm for Mining Association Rules in Large Databases. Proceedings of the 21st Conference on Very Large Data Bases. pp432-444. Zurich, Switzerland
- Spiliopoulou, M. (1999). The Laborious Way from Data Mining to Web Mining. International Journal of Comp. Sys., Science & Engineering. Special Issue on Semantics of the Web **14**:113-126
- Spiliopoulou, M., J.F. Roddick. (2000). Higher Order Mining: Modelling and Mining the Results of Knowledge Discovery. In: N. Ebecken, C.A. Brebbia. (eds.). Data Mining II. Proceedings of the Second International Conference on Data Mining Methods and Databases. pp309-320. WIT Press
- Thomas, S., S. Bodagala, K. Alsabti, S. Ranka. (1997). An Efficient Algorithm for the Incremental Updation of Association Rules in Large Databases. Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97). pp263-266. Newport Beach, California, USA
- Toivonen, H. (1996). Sampling Large Databases for Association Rules. Proceedings of the 22nd Conference on Very Large Data Bases. Mumbai (Bombay), India
- Wang, K., J. Tan. (1996). Incremental Discovery of Sequential Patterns. SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery
- Wang, K. (1997). Discovering Patterns from Large and Dynamic Sequential Data. Intelligent Information Systems **9**:8-33

6.4.6 Conclusions Technology trends

Jeroen Meij

Hardware

Several technical trends can be observed in the area of data mining.
In hardware there is a distinct trend towards distributed and parallel computing. SM-MIMD[1] architectures are popular because of the ease of modifying sequential data mining algorithms for this architecture. DM-MIMD[2] architectures get a lot of attention from data mining, but have several drawbacks that prevent wide scale application. A special type of SM-MIMD machine, ccNUMA[3] has a relatively small number of tightly integrated processors with shared memory properties. As I/O and operating systems improve, ccNUMA machines will play an important role in data mining applications. In the near future, the application of cheap Beowulf clusters development (of workstations or PC's) will rise as a result of increasing network speed, as for example defined in the Infiniband protocol. Field Programmable Gate Arrays (FPGAs) hardware may be configured to perform new tasks, achieving the optimal configuration for every operation. Finally, holographic storage could offer performance advantages for data mining by allowing for very fast page-wise correlation of data.

Parallel data mining

Data mining algorithms and underlying techniques can be parallelized to make them effective in the analysis of very large data sets. Several parallel strategies, algorithms, techniques and prototypes have been developed in the recent years. They allow researchers and end-users to mine large databases offering scalable performance.

Parallel execution of different data mining algorithms and techniques can be integrated to obtain a better model, not just to get high performance, but also high accuracy. Some promising research directions:

- Not just parallel algorithms, but environments and tools for interactive high performance data mining and knowledge discovery.
- Parallel text mining.
- Parallel and distributed web mining.
- Integration of parallel data mining with parallel data warehouses.
- Use of parallel computing in all the phases of the KDD process and support of efficient data warehouses.

An important research topic is the integrated use of clusters and grids for distributed and parallel knowledge discovery. The development of software architectures, environments and tools for grid-based data mining will result in Grid-

1 Shared Memory Multiple Instruction stream Multiple Data stream.

2 Distributed Memory MIMD.

3 Cache Coherent Non Uniform Memory Access.

aware PDKD[1] systems that will support high performance data mining applications on geographically distributed data sources.

Relational data mining

A new development is data mining on relational data, which is being extended to object oriented databases. Domain knowledge and distributed environments can be integrated in the data mining process by using the object oriented UML, Unified Modeling Language.

Pattern evolution

When data mining has revealed patterns from a database, an evolution in the patterns will occur when the data changes. These changes may be:

- Changes in the statistics of a pattern.
- Discovery of a new pattern.
- Changes in the patterns content.
- Disappearance of a pattern.

Updating knowledge or rules derived from changing data requires special strategies, to be divided in two types:

- incremental miners, i.e. algorithms that are capable of updating previously discovered rules as data change, most of them building on their static counterparts. They create an incremental update of previously discovered association rules.
- monitors that investigate when patterns or rules are affected by updates to the dataset. These algorithms often use existing mining algorithms, or they simply monitor the data rather than mining it.

When investigating temporal features, two approaches have emerged :

- the use of user-given constraints on the time domain to discover patterns belonging to the specified time interval,
- the use of given patterns to determine time intervals in which this rule holds.

A combined approach could be employed to discover patterns, along with their valid time intervals from scratch, providing a direction for future work.
A framework that is capable of dealing with all four types of changes a rule may undergo, is still missing. This might be a direction for future work.

1 Parallel and Distributed Knowledge Discovery.

Appendix CD-rom contents

Background

- The Moving Frontier: Archiving, Preservation and Tomorrow's Digital Heritage
  Hilary Berthon, Colin Webb, 10th VALA Biennial Conference and Exhibition, Melbourne, Victoria, 2000
  Paper, 2000
- Scenarios for Ambient Intelligence in 2010
  Compiled by K. Ducatel, M. Bogdanowicz, F.Scapolo, J. Leijten, J-C. Burgelman, IPTS-Seville, European Commission Community Research, User Friendly Information Society
  Final Report, 2001
- How Long before Superintelligence?
  Nick Bostrom, 1998, Department of Philosophy, Logic and Scientific Method London School of Economics
  Article, 1998
- Augmented Reality Survey
  Ronald T. Azuma, Hughes Research Laboratories 3011 Malibu Canyon Road, MS RL96 Malibu, CA 90265
  Survey Article, 1997

Tutorials, introductions and courses

- Electronic Statistics Textbook
  StatSoft, Inc. (1999). Tulsa, Oklahoma
  Electronic Textbook on Statistics
  http://www.statsoft.com/textbook/stathome.html
- An Introduction to Data Warehousing
  V.R. Gupta, 1997, System Services Corporation, Chicago, Illinois
  White paper
- Fuzzy and Neural Control
  Robert Babu_ka, 2000 Control Engineering Laboratory Faculty of Information Technology and Systems, Delft University of Technology, Delft, the Netherlands
- Symbolic Logic
  Peter Suber, 1996
  Philosophy Department, Earlham College
  Course Handout
- Logical Systems
  Peter Suber, 1999
  Philosophy Department, Earlham College
  Course Handout
- Probability and Statistics for Biological Sciences
  David W. Sabo, 2000
  British Columbia Institute of Technology, Canada, Course Math 2441
  Web Course
- Boyle, R., S. Hanlon, 2001
  The University of Leeds' Course on Hidden Markov Models, University of Leeds, United Kingdom
  http://www.scs.leeds.ac.uk/scs-only/teaching-materials/HiddenMarkovModels/html_dev/main.html
- Artificial Neural Networks
  Dr. Diego Andina de la Fuente, M. en I. Antonio Vega Corona, Francisco J. Rodríguez, 2001
  Universidad Politécnica de Madrid-UPM, España
  Web Tutorial
- Lecture Notes on Statistics and Data Analysis with Vista
  University of North Carolina, Forrest Young, 1999
  http://www.visualstats.org
- The Free On-line Dictionary of Computing
  http://www.foldoc.org, Editor Denis Howe
- Machine Learning Net Training Information Server
  Mlnet, 2001

Technical

Visualization

- Three Dimensional Information Visualisation
  Peter Young, Department of Computer Science University of Durham, Durham
  Technical Report, 1996
- Interactive Visualization of Large Graphs and Networks
  Tamara Munzner, Stanford University
  Ph D Dissertation, 2000
- Data Exploration Using Self-Organizing Maps
  Samuel Kaski
  Helsinki University of Technology
  Ph D Thesis, 1997
- Proximity Visualization of Abstract Data
  Wojciech Basalaj
  Dissertation, 2001

Extended articles

- Bibliometric Mapping Project of Mathematics and Computer Science and the Share of Swiss Research, carried out for the Swiss Science and Technology Council.
  A.F.J. van Raan, Noyons, E.C.M, CWTS, Leiden University
  Interactive Report, 1999, includes interactive versions of the maps in Section 2.2.3
- An Introduction to Data Warehousing
  V.R. Gupta
  System Services Corporation, Chicago, Illinois
  http://www.system-services.com
  White paper, 1997
- Rule Induction by Bump Hunting
  Ad Feelders, Utrecht University
  Extended version of Section 6.2.13 of this Book, 2001
- Association Rules
  Arno Siebes, Utrecht University
  Extended Version of Section 6.2.12 of this book, 2002
- Association Rules and Mechanizing Hypotheses Formation
  Jan Rauch, Laboratory of Intelligent Systems, Faculty of Informatics and Statistics, University of Economics, Prague
  Paper, 2001
- Guha and Kex for Knowledge Discovery in Economic Data
  Petr Berka, Jan Rauch, Laboratory of Intelligent Systems, Faculty of

Informatics and Statistics, University of Economics, Prague
Paper, 1997

- On Globally Robust Confidence Intervals for Regression Coefficients
  Matías Salibián-Barrera
  School of Mathematics and Statistics, Carleton University
  Thesis Proposal, 1998
- Data Fusion: A Way to Provide More Data to Mine in?
  Peter van der Putten, Sentient Machine Research
  Article, 1999
- Dewey Goes Surfing: Agent-Based Information Retrieval and Classification Support
  Bjørn Christian Tørrissen, Norwegian University of Science and Technology, Faculty for Physics, Informatics and Mathematics
  Master Thesis, 1998
- Rough Sets as a Framework for Data Mining
  Øyvind Tuseth Aasheim, Helge Grenager Solheim
  Knowledge Systems Group, Faculty of Computer Systems and Telematics
  The Norwegian University of Science and Technology, Trondheim
  Project Report, 1996
- Machine Learning on Non_Homogeneous, Distributed Text Data
  Dunja Mladenic, University of Ljubljana, Faculty of Computer and Information Science
  Ph D Thesis, 1998
- Tracking Conversational Context for Machine Mediation of Human Discourse
  MIT Media Laboratory, Perceptual Computing Technical Report #530
  Appeared in: AAAI Fall Symposium — Socially Intelligent Agents — The Human in the Loop, November 2000
  Tony Jebara, Yuri Ivanov, Ali Rahimi, Alex Pentland
  MIT Media Lab, 20 Ames St., Cambridge, MA 02139
  Technical Report, 2000
- Constructing Knowledge from Multivariate Spatiotemporal Data: Integrating Geographic Visualization (GVis) with Knowledge Discovery in Database (KDD) Methods
  Alan M. MacEachren, Monica Wachowicz, Robert Edsall, and Daniel Haug, Department of Geography, Penn State University, and Raymon Masters, Center for Academic Computing, Penn State University
  Paper, 1999
- Integrating GVis, GIS and KDD for Exploring Spatio-Temporal Data
  Monica Wachowicz, Wageningen UR, Centre for Geo-Information, The Netherlands
  Paper, 1999

– How Can Knowledge Discovery Methods Uncover Spatio-Temporal Patterns in Environmental Data?
  Monica Wachowicz, Wageningen UR, Centre for Geo-Information, The Netherlands
  Paper, 2000
– The Integration of Geographic Visualization with Knowledge Discovery in Databases and Geocomputation
  Mark Gahegan, Monica Wachowicz, Mark Harrower and Theresa-Marie Rhyne
  Paper, 2001
– GeoInsight: an Approach for Developing a Knowledge Construction Process Based on the Integration of GVis and KDD Methods
  Monica Wachowicz, Wageningen UR, Centre for Geo-Information, The Netherlands
  Paper, 2001
– Uncovering Spatio-Temporal Patterns in Environmental Data
  Monica Wachowicz, Wageningen UR, Centre for Geo-Information, The Netherlands
  Paper, 2002
– Three Companions for Data Mining in First Order Logic, © Springer Verlag
  Luc De Raedt, Hendrik Blockeel, Luc Dehaspe, Wim Van Laer, Department of Computer Science, Katholieke Universiteit Leuven, Belgium, from: Relational Data Mining (S. Džeroski, and N. Lavrač, eds.), 2001, pp105-139, with Kind Permission of Springer Verlag
  Chapter, 2001
– Inductive Logic Programming: Techniques and Applications, Nada Lavrač and Sašo Džeroski, Ellis Horwood, New York, 1994
  Updated Abstract, 2002; Book, 1994

Applications

– Credit Rating Prediction Using Self-Organizing Maps
  Roger P.G.H. Tan, Erasmus University Rotterdam, Faculty of Economics
  Master Thesis, 2000
– NASA Workshop on Issues in the Application of Data Mining to Scientific Data, October 19–21, 1999, University of Alabama in Huntsville, Huntsville, Alabama
  Final Report, 1999
– An Investigation into the Use of Maximum Likelihood Classifiers, Decision Trees, Neural Networks and Conditional Probabilistic Networks for Mapping and Predicting Salinity
  Fiona Evans, Curtin University of Technology, Australia
  Masters Thesis, 1998

– Adaptive Systems Management
Syllogic
White Paper, 1998
– Web Usage Mining and Discovery of Association Rules from HTTP Servers Logs, Gabriele Bartolini, Monash University, Melbourne, Victoria, Australia
Paper, 2001
– Approximate Inference for Medical Diagnosis
J.J. Wiegerinck , H.J. Kappen, E.W.M.T. ter Braak, W.J.P.P. ter Burg, M.J. Nijman, Y.L. O, and J.P. Neijt, Foundation for Neural Networks, Nijmegen University, University Medical Centre, Utrecht University, Academic Medical Centre, The Universiteit van Amsterdam, The Netherlands
Paper, 1999
– VICAR Video Navigator:Content Based Video Search Engines Become a Reality
Peter van der Putten, Sentient Machine Research
Article
– Neural Networks based Data Mining and Knowledge Discovery in Inventory Applications
Kanti Bansal, Sanjeev Vadhavkar, Amar Gupta
Paper, 1996

Standards and directives

– OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data, OECD. 1980
– Directive 94/EC of the European Parliament and of the Council on the Protection of Individuals With Regard to the Processing of Personal Data And on the Free Movement of Such Data. 1995
– MPEG-21 Overview
ISO/IEC JTC1/SC29/WG11/N4318
INTERNATIONAL ORGANISATION FOR STANDARDISATION, Requirements Group, Eds. Jan Bormans, Keith Hill. 2001
– MPEG-4 Overview — (V.18 — Singapore Version)
ISO/IEC JTC1/SC29/WG11 N4030
INTERNATIONAL ORGANISATION FOR STANDARDISATION, WG11 (MPEG), Editor: Rob Koenen
Final Report, 2001
– Overview of the MPEG-7 Standard (Version 4.0)
ISO/IEC JTC1/SC29/WG11 N3752
INTERNATIONAL ORGANISATION FOR STANDARDISATION, Requirements, Editor: José M. Martínez (UPM-GTI, ES). 2000
– Extensible Markup Language (XML) 1_0 (Second Edition). 2000

Software on the CD-rom

The STT 65 CD-rom contains software that provides the reader the opportunity to experiment with many of the techniques described in the book. *Although most software included is free to use for testing, educational or non commercial purposes, for the licensing terms and conditions we refer to the statements included with the software packages. Restrictions may vary for each application. All executable files are for windows 98, unless stated otherwise. Some packages may be available for other platforms from the home pages of the authors.

software	**Perot systems Texthub**	
related section	5.4.2 Text mining,	
	6.2.20 Text mining techniques	
topics	Text mining, search and retrieval	
license	Demo limited to mining the STT 65 CD-rom contents: full text of this book and over 50 additional documents	
path	See interface (index.htm) on CD-rom	
URL	http://www.perotsystems.nl	

software package	**ADE-4**	
related section	6.2.2 Regression,	
	6.2.3 Discriminant analysis,	
	6.2.4 Subspace methods	
topics	Principal Component Analysis, correspondance analysis, spatial data analysis, discriminant analysis, linear regression analysis, graphical display techniques, multivariate graphics	
license	Free*	
path	Software\ pca_da_regr\	
file(s)	ADE4PC.exe	*user interface*
	ADEWin95-213.84.180.149.tar	*archive of modules*
documentation	DownloadADE-4.htm	*info*
	What is ADE-4.htm	
literature	install.pdf	
	interface.pdf	
URL	http://pbil.univ-lyon1.fr/ADE-4/ADE-4.html	

software package	**BNPC** **– Bayesian network powerconstructor** **– powerpredictor** **– data preprocessor**	
related section	6.2.11 Belief networks/Bayesian networks,	
	6.2.8 Naïve Bayes classifier	
topics	construct Bayesian networks from data, modeling, classification and prediction, data preprocessing	
license	Free* (windows 98, NT)	
path	software\Bayesian_class\	
file(s)	readme.htm (installation instructions)	
	instmsia.exe (win 98)	
	instmsiw.exe (win NT)	
documentation	introduction	
	BNPC\bnsoft.htm	
	user manual, tutorial	
	BNPC\bnpchlp\index.html	
literature	papers and technical reports:	
	BNPC\Doc\aistat97.pdf	An Algorithm for Bayesian Belief Network Construction from Data
	BNPC\Doc\cikm97.pdf	Learning Belief Networks from Data: An Information Theory Based Approach
	BNPC\Doc\reportoo.pdf	Learning Bayesian Belief Network Classifiers: Algorithms and System
URL	http://www.cs.ualberta.ca/~jcheng/bnsoft.htm	

software package	**Bayesbuilder**	
related section	2.3.2 Decision support for medical diagnosis,	
	6.2.11 Belief networks/Bayesian networks	
topics	Bayesian networks	
license	Free*	
path	software\Bayesian_net	
file(s)	BayesBuilder1.1.9.zip	*java app.*
	j2re-1_3_1_02-win-i.exe	*java runtime 1.3*
documentation	Bayesbuilder.htm	
literature	See Section 6.2.11	
remarks	First install java runtime 1.3	
URL	http://www.snn.kun.nl/nijmegen/index.php3?page=31	

software package	CLUTO
related section	Clustering
topics	Clustering high dimensional datasets, analyzing the characteristics of the various clusters. Partitional and agglomerative clustering algorithms
license	Free*
path	software\clustering\
file(s)	cluto-1.5.1.zip
documentation	CLUTO.htm
remarks	http://www-users.cs.umn.edu/~karypis/cluto/index.html

software package	Genetic-Adapt FuzzyWare (GAF)	
related section	6.2.15 Fuzzy logic techniques	
topics	Fuzzy logic	
license	Free* (ms dos)	
path	software\fuzzy\	
file(s)	Readme.txt (installation instructions)	
	gaf2oob.zip	
documentation	gaf.doc	*the user guide*
	fcl.doc	*Fuzzy Control Language guide*
	refer.doc	*book and papers for further references*
	demo.doc	*document for demonstration*
	tutorial.doc	*tutorial for building GAF application*
	abstract.doc	*an abstract of GA FuzzyWare*
	intro.doc	*an introduction to GA FuzzyWare*
URL	http://www.eps.ufsc.br/~martins/fuzzy/fuzzy/huang.txt	

software package	LISp-Miner	
related section	2.2.4 Mining for scientific hypotheses,	
	6.2.12 Association rules,	
	6.2.7 Classification	
topics	Association rules, implication and equivalency association rules, hypothesis tests, conditional association rules	
license	Free*, academic	
path	software\association_rules\	
file(s)	index.htm info	
	download\files\LISp_Miner.zip	*The LISp-Miner system*
	download\files\Barbora.zip	*Demo based on data of the fictional bank Barbora*

	download\files\help.zip	*help files for LISp-Miner*
documentation	index.htm info	
literature	See Section 2.2.4	
remarks	For updates and additional packages see the website	
URL	http://lispminer.vse.cz/	

software package	**Neural Networks at your fingertips, Karsten Kutza**	
related section	6.2.7 Neural networks	
topics	Neural networks: Adaline network, Backpropagation, Hopfield model, Bi-directive associative memory (BAM), Boltzmann machine, Counter propagation network (CPN), Self organizing map (SOM), Adaptive resonance theory (ART1)	
license	Free*	
path	software\neural_nets\suite	
file(s)	index.html	*introduction*
	nn.zip	*sources and win 32 executable files*
documentation		
literature		
remarks	DIY kit of ANSI C sources for those who want to program neural nets	
URL	http://www.geocities.com/CapeCanaveral/1624/	

software package	**PERMAP 8.0**
related section	6.2.4 Multidimensional scaling
topics	Multidimensional scaling, visualization. PERMAP is an interactive computer program for making MDS analyses. It offers both metric and non-metric MDS techniques. It is PC-based and visually oriented
license	Free*
path	software\md_scaling\
file(s)	Permap.zip
documentation	PermapManual.doc (manual and installation instructions)
literature	see PermapManual.doc
URL	http://www.ucs.ull.edu/~rbh8900/permap.html

software package	Pittnet Neural Network Educational Software	
related section	6.2.7 Neural networks	
topics	Neural networks: backpropagation (BP), Kohonen self-organizing, adaptive resonance theory I (ART I), and radial basis function (RBF)	
license	Free*	
path	software\neural_nets\pittnet\	
file(s)	readme.txt (installation instructions)	
	pittnet.exe	
documentation	Guide.pdf	*user guide*
URL	http://www.pitt.edu/~aesmith/	

software package	ROSETTA	
related section	6.2.16 Rough sets	
topics	Rough sets	
license	Free*	
path	software\rough_sets	
file(s)	setup.txt	*installation instructions*
	rosettasoftware.zip	*win32*
	The ROSETTA Homepage.htm	*introduction*
	features.htm	*features*
	Download.htm	*additional files*
documentation	manual.pdf	*user manual*
literature	Discernibility and Rough Sets in Medicine: Tools and Applications PHD thesis of Aleksander Øhrn (1999)	
URL	http://www.idi.ntnu.no/~aleks/rosetta/	

software package	Synope summarizer
related section	Text mining Section 5.4.1
topics	Automatic summarizing software — IE browser add on
license	30 days Demo (English)
path	software\synope\
file(s)	setup.exe
URL	http://www.carp-technologies.nl

software package	**Vista** **Log-lin plug in** **Multivariate plug in**	
related section	6.2.2 Regression,	
	6.2.4 Subspace methods,	
	6.2.5 Multidimensional scaling	
topics	Interactive data visualization: Spinplots, Scatterplots, Scatterplot Matrices, Histograms, Boxplots, Parallel Coordinate Plots, Mosaic Plots, Quantile Plots, Normal Probability Plots, Quantile-Quantile Plots, Diamond Plots, Dotplots, Biplots, and Guided Tour Plots.	
	Statisitics: Means, Standard Deviations, Variances, Ranges, Quartiles, Medians, Correlations, Covariances, Distances, Frequency Tables.	
	Univariate Analysis: Univariate Tests including T- and Z-tests (confidence intervals) ANOVA, Multiple Regression.	
	Multivariate Analysis: Multiple Regression, Principal Component Analysis, Multidimensional Scaling, Correspondence Analysis.	
license	Free*	
path	software\statistical_methods	
file(s)	ViSta64-Installer.exe	*Main program, win 32*
	Loglin64-Installer.exe	*Plug in*
	MulVar64-Installer.exe	*Plug in*
documentation	\papers\Forrest.psych\class-notes.html	
	Lecture Notes on Statistics and Data Analysis with Vista.	
	User manual documents	
	front.pdf	*Frontmatter*
	chapo1.pdf	*Chapter 1:Introduction f*
	chapo2.pdf	*Chapter 2:Tutorial*
	chapo3.pdf	*Chapter 3:Defining Data, References*
	chapo8.pdf	*Univariate Multiple Regression*
	anova.pdf	*Analysis of Variance*
	pca.pdf	*Principal Components Analysis*
	chap11.pdf	*Correspondence Analysis*
	devel.pdf	*Enhancing ViSta, Developing Statistical Objects*
	The lecture notes describe statistical concepts and uses vista for examples.	
literature	See www.visualstats.org	
URL	http://www.visualstats.org	

software package	Weka	
related section	6.2.2 Regression,	
	6.2.17 Support vector machines,	
	6.2.7 Neural networks,	
	6.2.8 Naïve Bayes classifiers,	
	6.2.11 Association rules,	
	6.2.18 Combining classifiers	
topics	Clustering:	
	Cobweb and an EM algorithm	
	Classification:	
	decision tree inducers	
	rule learners	
	naive Bayes	
	decision tables	
	locally weighted regression	
	support vector machines	
	instance-based learners	
	logistic regression	
	voted perceptrons	
	multi-layer perceptron	
	Numeric prediction:	
	linear regression	
	model tree generators	
	locally weighted regression	
	instance-based learners	
	decision tables	
	multi-layer perception	
	Implemented 'meta-schemes' include:	
	bagging	
	stacking	
	boosting	
	regression via classification	
	classification via regression	
	cost sensitive classification	
	Association rule learner:	
	Apriori	
license	Free*	
path	software\weka_datm_suite	
file(s)	Weka-3-2-1jre.exe	*win 98 install file including java*
	run time	
	datasets-UCI.jar	*datasets in compressed format*
	datasets-numeric.jar	*numerical datasets*

documentation	Experiments.pdf
	Weka 3 — Data Mining with Open Source Machine Learning
	Software in Java.htm *introduction*
	The manual is included in the weka install file
literature	See http://www.cs.waikato.ac.nz/~ml/weka/index.html
URL	http://www.cs.waikato.ac.nz/~ml/weka/index.html

software package	**Metal for Weka**
related section	Section 6.1.5 Process embedding, 6.4.4 Meta learning
topics	Meta learning for technique selection
license	GNU*
path	software\metal\
file(s)	Wekametal.htm info
	WekaMetal\WekaMetal.jar
documentation	Wekametal.htm info
literature	See Section 6.4.4
remarks	Requires Weka 3.2
URL	http://www.cs.bris.ac.uk/~farrand/wekametal/

Survey organization

This publication was realized with the active help of 93 data mining experts. STT wishes to express its gratitude to all those who donated their time and energy to this study project. A steering committee was set up to guard the academic content of this study project.

Steering Committee

P.W. Adriaans	Perot Systems Netherlands BV, Amersfoort
	Universiteit van Amsterdam
A.E. Eiben	Vrije Universiteit Amsterdam
J.H.A.M. Grijpink	Ministry of Justice, The Hague
H.J. van den Herik	Universiteit Maastricht
J.N. Kok	Leiden University, LIACS
A.Y.L. Kwee	New Business Associates, Abcoude
(chairman taskforce)	
J.H. van der Veen	STT, The Hague
A.F.J. van Raan *(chairman)*	Leiden University, CWTS
A.P.J.M. Siebes	Utrecht University
(chairman taskforce)	
M.J. den Uyl	Sentient Machine Research B.V., Amsterdam
B.J. Wielinga	Universiteit van Amsterdam

Project participants

J. van den Berg	Erasmus University Rotterdam
G. Beijer	Nedstat BV, Diemen
A. van den Bosch	Tilburg University
M.C. Bouvy	Bolesian, Utrecht
A.J. Feelders	Utrecht University
A.V. Groenink	Eidetica, Amsterdam
A. Hanjalić	Delft University of Technology
W. van der Hoek	Utrecht University
W.K. Hofland	IQUIP Informatica B.V., Diemen
G.J. Houben	Eindhoven University of Technology
W.A. Kosters	Leiden University, LIACS, Leiden
J.J. Meulman	Leiden University
M. Poel	Twente University
H. La Poutré	CWI, National Research Centre for Mathematics
	and Computer Science, Amsterdam
P. van der Putten	Leiden University, LIACS, Leiden
M.J.T. Reinders	Delft University of Technology
E. Schreuders	Tilburg University
P.M.A. Sloot	Universiteit van Amsterdam
M. van Someren	Universiteit van Amsterdam
J. Veenstra	Tilburg University
F. Verdenius	ATO, Wageningen
G. Vriend	Nijmegen University

J. van der Wal	Dutch National Police Agency, Driebergen Rijsenburg
L.F.A. Wessels	Delft University of Technology
J. Wittmaekers	CV-ROM, Groningen

Project authors

P.W. Adriaans	Perot Systems Netherlands BV, Amersfoort, Universiteit van Amsterdam
J. van den Berg	Erasmus University Rotterdam
A. van den Bosch	Tilburg University
M.C. Bouvy	Bolesian, Utrecht
A.E. Eiben	Vrije Universiteit Amsterdam
A.J. Feelders	Utrecht University
J.H.A.M. Grijpink	Ministry of Justice, The Hague
A. Hanjalić	Delft University of Technology
H.J. van den Herik	Universiteit Maastricht
W. van der Hoek	Utrecht University
G.J. Houben	Eindhoven University of Technology
W.A. Kosters	Leiden University, LIACS, Leiden
A.Y.L. Kwee	New Business Associates, Abcoude
J.J. Meulman	Leiden University
H. La Poutré	CWI, National Research Centre for Mathematics and Computer Science, Amsterdam
P. van der Putten	Leiden University, LIACS, Leiden
A.F.J. van Raan	Leiden University, CWTS
M.J.T. Reinders	Delft University of Technology
A. Siebes	Utrecht University
P.M.A. Sloot	Universiteit van Amsterdam
M. van Someren	Universiteit van Amsterdam
F. Verdenius	ATO, Wageningen
J. van der Wal	Dutch National Police Agency, Driebergen Rijsenburg
L.F.A. Wessels	Delft University of Technology

External authors

R. Babuška	Delft University of Technology
S. Baron	Humboldt University Berlin, Germany
C.T.M. Baten	Roessingh Research and Development, Enschede
R.G. Belleman	Universiteit van Amsterdam

W.-M. van den Bergh	Erasmus University Rotterdam
P.D. Bezemer	Vrije Universiteit Amsterdam
H. Blockeel	Katholieke Universiteit Leuven, Leuven, Belgium
R. Boyle	University of Leeds, United Kingdom
E.W.M.T. ter Braak	Utrecht University
A. Cavoukian	Information and Privacy Commissioner, Ontario, Canada
N. Brandt	Perot Systems Netherlands BV, Amersfoort
L. Dehaspe	PharmaDM, Heverlee, Belgium
E.M.L. Dusseldorp	Leiden University
T. Gevers	Universiteit van Amsterdam
C. Giraud-Carrier	ELCA Informatique SA, Lausanne, Switzerland
P. Goethals	Ghent University, Gent, Belgium
Q. Grens	Perot Systems Netherlands BV, Amersfoort
M. de Haas	Perot Systems Netherlands BV, Amersfoort
D.J. Hand	Imperial College, London
R.M. van Hees	Space Research Organisation Netherlands (SRON), Utrecht
T. Heskes	SNN, University of Nijmegen
C.J Huberty	University of Georgia, Athens, USA
H.J. Kappen	Nijmegen University
J. Keller	Daimler Chrysler AG, Ulm, Germany
R.E.J. Keller	Leiden University, LIACS, Leiden
R. Kosala	Katholieke Universiteit Leuven, Leuven, Belgium
J. Krikken	National Museum of Natural History, Leiden
W. Kowalczyk	Vrije Universiteit Amsterdam
M. Leman	Ghent University, Ghent, Belgium
R.C. van Lent	IBM Global Services, Uithoorn
D.H. Lie	Carp technologies, Hengelo
F. Neven	University of Limburg (LUC), Diepenbeek, Belgium
J.C. Noordam	ATO, Wageningen
T.C. Noordermeer	National Library of The Netherlands, The Hague
E.C.M. Noyons	Leiden University, CWTS
E. Pekalska	Delft University of Technology
W.H. Piel	National Museum of Natural History, Leiden
P.P.J. Ramaekers	Tiaram B.V., Weert
J. Rauch	University of Economics, Prague
D. de Ridder	Delft University of Technology

P.M. van Rosmalen	Aurus Knowledge &Training Systems BV, Maastricht
M. Schuemie	Delft University of Technology
H.G. Solheim	Computas AS, Lysaker, Norway
W.J. Som de Cerff	Royal Netherlands Meteorological Institute (KNMI), De Bilt
M. Spiliopoulou	Handelshochschule Leipzig (HHL), Germany
A.J. van der Steen	Utrecht University
D. Talia	ISI-CNR, Institute of System Analysis and Information Technology, Italy
R.P.G.H. Tan	Robeco Group N.V, Rotterdam
M. Trautwein	Perot Systems Netherlands BV, Amersfoort
S. Tsumoto	Shimane Medical University, Enya-cho Izumo City, Japan
J. van de Vegte	Royal Netherlands Meteorological Institute (KNMI), De Bilt
J. Verbeek	Universiteit Maastricht
M. Wachowitz	Wageningen UR
L. van Wel	Eindhoven University of Technology
C.R. Westphal	Visual Analytics Incorporated, Poolesville, USA
W.A.J.J. Wiegerinck	Nijmegen University
O.R. Zaïane	University of Alberta, Alberta, Canada
D.L.T. Zwietering	IBM Global Services, Uithoorn

Project management

The study project was managed by Jeroen Meij, project manager STT. Rosemarijke Otten, project secretary at STT, assisted him in the organization of the study project. The discussions and advice of the former and present directors of STT, Erik van de Linde and Hans van der Veen, also contributed to the project. Rosemarijke Otten participated in the editing of the publication and took care of the linguistic editing.

STT Publications

All publications with an ISBN number can be ordered from the Netherlands Study Centre for Technology Trends (STT/Beweton) or from the book shop.

The remaining publications are only available from:
The Netherlands Study Centre for Technology Trends (STT/Beweton)
P.O. Box 30424
2500 GK The Hague
The Netherlands
Telephone +31 70 3029830
Fax +31 70 3616185
E-mail info@stt.nl

The most recent list of publications can be found on the homepage:
http://www.stt.nl

65 Dealing with the data flood, Mining data, text and multimedia
edited by J.M. Meij, 2002 (ISBN 90-804496-6-0)

64 Reliability of technical systems, anticipating trends
edited by M.R. de Graef, 2001 (ISBN 9084496 5 2)

63 Future@work.nl, reflections on economy, technology and work
edited by Rifka M. Weehuizen, 2000 (ISBN 9084496 4 4)

62 Innovation in product development, strategy for the future
edited by Arie Korbijn, 1999 (ISBN 90 804496 3 6)

61 Rapid current, the next electrical innovation wave
edited by J.M. Meij, 1999 (ISBN 90 804496 2 8)

60* Nanotechnology, towards a molecular construction kit
Edited by Arthur ten Wolde, 1998 (ISBN 90 804496 1 X)

59 Buildingwise; materials and methods for future buildings
edited by Annemieke Venemans, 1997 (ISBN 90 61 55 816 6)

58 Healthy production; innovating for improved working conditions
edited by Arie Korbijn, 1996 (ISBN 90 61 55 7445)

57 Digital tools for vocational training
edited by Arthur ten Wolde, 1996 (ISBN 90 61 55 7305)

56* Microsystem technology: exploring opportunities
Edited by Gerben Klein Lebbink, 1994 (ISBN 90 14 05088 7)

55 Clean opportunities: ideas on entrepreneurship and environmental management
edited by E.W.L. Van Engelen and J. van Goor, 1994 (ISBN 90 04929 3)

54 Short-haul freight transport
edited by M.J. Venemans, 1994 (ISBN 90 14 04928 5)

53 Electricity in perspective: 'energy and environment' in industry
edited by E.W.L. Van Engelen, 1992 (ISBN 90 14 04715 0)

52 Dealing with complexity
edited by M.J.A. Alkemade, 1992 (ISBN 90 14 03883 6)

51 Agricultural commodities for industry
edited by W.G.J. Brouwer, 1991 (ISBN 90 14 03882 8)

50 Vocational training for the future: instrument for policy
edited by H.B. Van Terwisga and E. van Sluijs, 1990 (ISBN 90 14 04506 9)

49 Limits to technology
edited by A.J. Van Griethuysen, 1989 (ISBN 90 14 03880 1)

48 Expert systems in the manufacturing industry
J.J.S.C. De Witte and A.Y.L. Kwee, 1988 (ISBN 90 14 03758 9)

47 Expert systems in the service industry
edited by A.Y.L. Kwee and J.J.S.C. De Witte, 1987 (ISBN 90 14 03719 8)

46 Expert systems in medical decision-making
J.J.S.C. De Witte and A.Y.L. Kwee, 1987 (ISBN 90 14 03718 X)

* Available in English, the remainder in Dutch only.

45 Expert systems in education
edited by J.J.S.C. De Witte and A.Y.L. Kwee, 1987 (ISBN 90 14 03717 1)

44 Designing for maintenance, now and in the future
edited by G. Laurentius, 1987 (ISBN 90 14 03716 3)

43 New applications of materials
edited by A.J. Van Griethuysen, 1986

42 Engineering for the elderly
edited by M.H. Blom-Fuhri Snethlage, 1986 (ISBN 90 14 03822 4)

41 The future of our foodstuff industry
edited by J.C.M. Schogt and W.J. Beek, 1985 (ISBN 90 14 03821 6)

40 Industry, knowledge and innovation
edited by H. Timmerman, 1985 (ISBN 90 14 03820 8)

39 The vulnerability of the city; interruptions to water, gas, electricity and telecommunications
edited by G. Laurentius, 1984 (ISBN 90 6275 145 8)

38* Man and information technology: towards friendlier systems
edited by J.H.F. Van Apeldoorn, 1983 (ISBN 90 6275 136 9)

37 The Netherlands and the bounty of the sea: industrial perspectives and the new law of the sea
edited by J.F.P. Schönfeld and P.J. De Koning Gans, 1983 (ISBN 90 6275 111 3)

36 Information technology in the office; experiences in seven organizations
compiled by F.J.G. Fransen, 1983 (ISBN 90 6275 135 0)

35 Automation in the factory; directions for policy-making
edited by H. Timmerman, 1983 (ISBN 90 6275 112 1)

34 Flexible automation in the Netherlands; experiences and opinions
edited by G. Laurentius, H. Timmerman and A.A.M. Vermeulen, 1982

33 Future heating of homes and other buildings
edited by A.C. Sjoerdsma, 1982 (ISBN 90 6275 094 X)

32 Micro-electronics for our future; a critical appraisal
compiled by Viscount E. Davignon et al., 1982 (ISBN 90 6275 089 3)

31-9 Micro-electronics in the inland revenue office

31-8 Micro-electronics in the travel industry (ISBN 90 6275 073 7)

31-7 Micro-electronics in the office

31-6 Micro-electronics in banking (ISBN 90 6275 071 0)

31-5 Micro-electronics and the design process (ISBN 90 6275 070 2)

31-4 Micro-electronics and innovation in consumer products and services for use in the home (ISBN 90 6275 069 9)

31-3 Micro-electronics and process innovation in electrometallurgy (ISBN 90 6275 068 0)

31-2 Micro-electronics in printing and publishing (ISBN 906275 067 2)

31-1 Micro-electronics in cattle farming (ISBN 90 6275 066 4)

* Available in English, the remainder in Dutch only.

31 Micro-electronics in business and industry: current position and future prospects
compiled by H.K. Boswijk, 1981 (ISBN 90 6275 064 8)

30* Biotechnology; a Dutch perspective
edited by J.H.F. Van Apeldoorn, 1981 (ISBN 90 6275 051 6)

29 Home and technology: yesterday's experience, ideas for tomorrow
edited by J. Overeem and G.H. Jansen, 1981 (ISBN 90 6275 053 2)

28 The distribution of consumer goods; information and communication in perspective
edited by R.G.F. De Groot, 1980 (ISBN 90 6275 052 4)

27 Coal for our future
edited by A.C. Sjoerdsma, 1980

26 Forests and timber for our future
edited by T.K. De Haas, J.H.F. Van Apeldoorn and A.C. Sjoerdsma, 1979

25 Data processing in the medical profession
edited by R.G.F. De Groot, 1979

24 Trends in industry
P. De Wolff et al., 1978

23 Industry in the Netherlands: a survey of problems and options
edited by H.K. Boswijk and R.G.F. De Groot, 1978

22 Materials for our society
edited by J.A. Over, 1976

21 New approaches to urban traffic and transport
edited by J. Overeem, 1976

20 Food for all; place and role of the EC
J. Tinbergen et al., 1976

19* Energy conservation: ways and means
edited by J.A. Over and A.C. Sjoerdsma, 1974

18 Man and the environment: cycles of matter
Stuurgroep en Werkgroepen voor Milieuzorg, 1973

17 Man and the environment: towards clean air
Stuurgroep en Werkgroepen voor Milieuzorg, 1973

16 Man and the environment: controlled growth
Stuurgroep en Werkgroepen voor Milieuzorg, 1973

15 Technological forecasting: methods and possibilities
A. Van der Lee et al., 1973

14 Technology and preventive medical examination
M.J. Hartgerink et al., 1973

13 Communication city 1985: electronic communication with home and business
J.L. Bordewijk et al., 1973

* Available in English, the remainder in Dutch only.

12 Electricity in our future energy supply: options and implications
H. Hoog et al., 1972

11 Transmission systems for electrical energy in the Netherlands
J.J. Went et al., 1972

10* Barge carriers: some technical, economic and legal aspects
W. Cordia et al., 1972

9 Nutrition in the Netherlands, now and in the future
M.J.L. Dols et al., 1971

8 Man and the environment: priorities and choices
L. Schepers et al., 1971

7* Electrical energy needs and environmental problems, now and in the future
J.H. Bakker et al., 1971

6 Cheap electrical energy and its impact on technological development in the Netherlands
P.J. Van Duin, 1971

5 Transitional procedures in transport
J.L.A. Cuperus et al., 1969

4 How to formulate medium-term planning policy
P.H. Bosboom, 1969

3 Means of transport
J.L.A. Cuperus et al., 1968

2 Technology and the shape of the future; a telescopic view of telecommunication
R.M.M. Oberman, 1968

1 Trends in technology and engineering
J. Smit, 1968

* Available in English, the remainder in Dutch only.

Other publications:

- Technology pushes limits, if you know what I mean
  STT/Toonder, 1997
- * New applications of materials
  edited by A.J. Van Griethuysen, 1988 (ISBN 0 9513623 0 5)
- Marine developments in the United States, Japan, France, Federal Republic of Germany, United Kingdom and the Netherlands: organisation, spheres of interest and budgets
  edited by J.F.P. Schönfeld and Ph.J. De Koning Gans, 1984
  (published by: Distributiecentrum Overheidspublikaties, The Hague, the Netherlands)
- The importance of STT
  Th. Quené, 1983
- Innovation, a new direction
  H.K. Boswijk, J.G. Wissema and W.C.L. Zegveld, 1980

* Available in English, the remainder in Dutch only.

This study came about thanks to the financial support from companies, government and the Royal Institution of Engineers (KIVI).

Financial support STT

Akzo Nobel
Arcadis
Bank Nederlandse Gemeenten
CMG Nederland
Commissie van Overleg Sectorraden
Corus Group
Cosun
CSM
Delft Instruments
DHV Beheer
Dow Benelux
DSM
Eldim
EnergieNed
Energieproduktiebedrijf UNA
Fugro
Gamma Holding
Haskoning Nederland
Heineken Nederland
Holland Railconsult
Hollandsche Beton Groep
ING Bank
InnovatieNetwerk

IQUIP Informatica
KEMA
KIVI
Koninklijke KPN
Koninklijke Schelde Groep
Koninklijke Ten Cate
Lucent Technologies
Micro*Montage
Ministerie van Economische Zaken
Ministerie van Landbouw, Natuurbeheer en Visserij
Ministerie van Onderwijs, Cultuur en Wetenschappen
Nederlandsche Apparatenfabriek Nedap
Nederlandse Gasunie
Nederlandse Unilever Bedrijven
NIB Capital
Océ-Technologies
Philips Electronics
Rabobank Nederland
PinkRoccade
Schneider MGTE
Sdu
Shell Nederland
Siemens Nederland
Simac Techniek
Solvay Nederland
Stichting Energieonderzoek Centrum Nederland
Stork
TBI Holdings
TNO
TNT Post Groep
Urenco
VNU
Vopak Oil Logistics Europe & Middle East
Vredestein

Sponsors for this publication

The publication was sponsored by VNU Publishers and by Perot Systems Netherlands BV

perotsystems is a proud sponsor of the

STT Publication and CD-rom

Dealing With the Data Flood

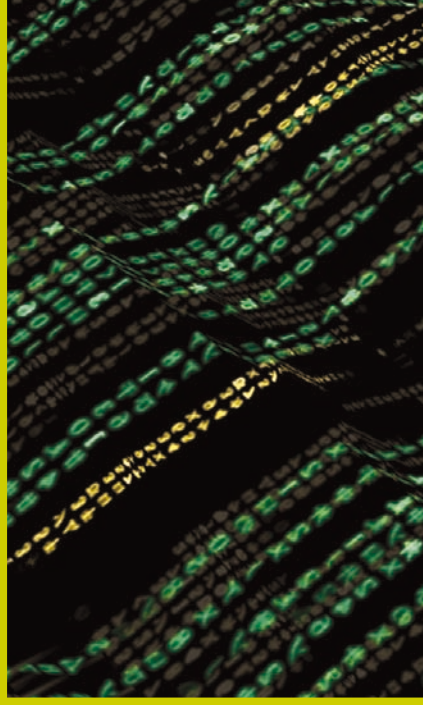
perotsystems®

IT. Services. Consulting. Solutions.

Perot Systems Netherlands BV
PO Box 2729
3800 GG Amersfoort
Hoefseweg 1
3821AE Amersfoort
The Netherlands
+31-33-4534545
www.perotsystems.nl

Perot Systems Incorporated
Global Headquarters
2300 West Plano Parkway
Plano, Texas 75075
+1-972-577-0000
www.perotsystems.com

STT Netherlands
Study Centre for
Technology Trends



In nearly every area of business and science – even in our private life – we are confronted with an increase in data flows. Data that is collected and processed. Data that holds valuable information, and may provide us knowledge, but often is inaccessible because of its form and volume.

Fortunately, new methods are emerging and evolving that enable us to create knowledge from data. SPECIFICALLY, METHODS AND TOOLS THAT EXTRACT PREVIOUSLY UNKNOWN INFORMATION FROM AGGREGATIONS OF DATA. *To name a few examples, patterns and relations can be revealed, clusters identified and predictions made. The application extends to many types of data: concepts and relations can be derived from large text collections, image and video collections can be analyzed for better access.*

This book gives an idea of the possibilities and expectations of data mining from five perspectives:

Science – The role of data mining for science in general is discussed and many examples are given.

Business and government – Starting from several common needs in these environments, over ten cases are presented illustrating present and future possibilities for data mining.

Ethics, privacy and legislation – Ethical aspects of web mining, fair information practices and some general legal aspects are examined.

The individual – From the perspective of the individual (one moment a knowledge worker, the next a private person) looking for an exciting scene in a sports match video a different view arises: text mining, web mining, image, video and music mining are the topics here. We also discuss the role of data mining in personal knowledge management.

Technology and techniques – After a discussion on methodology and process embedding, we will zoom in on twenty knowledge discovery, data mining and analysis techniques, the basic building blocks enabling us to convert data into knowledge.

The book is enhanced by a CD-rom, containing the full book text and many tutorials, reports, and papers both deepening and widening the view. All textual data is accessible through an integrated text mining tool.

Many of the techniques from the technical part of this book can be explored through freeware and demo software included on the CD-rom.

Sponsored by

perotsystems

VNU Publishers

ISBN 90-804496-6-0

Laser Proof

9 789080 449664