



### 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 than 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 marketers, 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<sup>1</sup> — 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.*

---

<sup>1</sup> 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.

