2

Data Warehousing, Data Mining, and OLAP

Objectives:

- This deals with the concept of data mining, need and opportunities, trends and challenges, data mining process, common and new applications of data mining, data warehousing, and OLAP concepts.
- It gives an introduction to data mining: what it is, why it is important, and how it can be used to provide increased understanding of critical relationships in rapidly expanding corporate data warehouse.
- Data mining and knowledge discovery are emerging as a new discipline with important applications in science, engineering, health care, education, and business.
- New disciplined approaches to data warehousing and mining are emerging as part of the vertical solutions approach.
- Extracting the information and knowledge in the form of new relationships, patterns, or clusters for decision making purposes.
- We briefly describe some success stories involving data mining and knowledge discovery.
- We describe five external trends that promise to have a fundamental impact on data mining.
- The research challenges are divided into five broad areas: A) improving the scalability of data mining algorithms, B) mining nonvector data, C) mining distributed data, D) improving the ease of use of the data mining systems and environments, and E) privacy and security issues for data mining.
- We present the concept of data mining and aim at providing an understanding of the overall process and tools involved: how the process turns out, what can be done with it, what are the main techniques behind it, and which are the operational aspects.
- OLAP servers logically organize data in multiple dimensions, which allows users to quickly and easily analyze complex data relationships.

S. Sumathi and S.N. Sivanandam: *Data Warehousing, Data Mining, and Olap*, Studies in Computational Intelligence (SCI) **29**, 21–73 (2006)
www.springerlink.com

- OLAP database servers support common analytical operations, including consolidation, drill-down, and slicing and dicing.
- OLAP servers are very efficient when storing and processing multidimensional data.

Abstract. This deals with the concept of data mining, need and opportunities, trends and challenges, process, common and new applications, data warehousing, and OLAP concepts. Data mining is also a promising computational paradigm that enhances traditional approaches to discovery and increases the opportunities for breakthroughs in the understanding of complex physical and biological systems. Researchers from many intellectual communities have much to contribute to this field. Data mining refers to the act of extracting patterns or models from data. The rate growth of disk storage and the gap between Moore's law and storage law growth trends represent a very interesting pattern in the state of technology evolution. The ability to capture and store data has produced a phenomenon we call the *data tombs* or *data stores* that are effectively write-only.

"Data Mining" (DM) is a folkloric denomination of a complex activity, which aims at extracting synthesized and previously unknown information from large databases. It also denotes a multidisciplinary field of research and development of algorithms and software environments to support this activity in the context of real-life problems where often huge amounts of data are available for mining. There is a lot of publicity in this field and also different ways to see the things. Hence, depending on the viewpoints, DM is sometimes considered as just a step in a broader overall process called *Knowledge Discovery in Databases* (KDD), or as a synonym of the latter as we do in this section. Thus, according to this purist definition DM software includes tools of automatic learning from data, such as machine learning and artificial neural networks, plus the traditional approaches to data analysis such as query and reporting, on-line analytical processing or relational calculus, so as to deliver the maximum benefit from data.

Data warehousing is a term that has been coined in recent years to describe computer systems designed to manage data for analysis and to assist management in decision making. A data warehouse can be simply described as "a data infrastructure specifically designed for query, analysis and reporting." While much of the data contained in these systems will be copied from administration systems, the data warehouse should also contain relevant external data, such as exchange rates, yield curves, and actuarial reserves. A question often asked is "why make the expense and effort of keeping copies of data that exists in other systems?" Many people can see the value in making all your data available from a single source, but the real answer is primarily to do with structuring the data such that it is most suitable for analysis.

Administration systems are designed for entering and retrieving data on individual persons or policies; they are not well suited for direct analysis. Also, the reality in many insurance companies is that duplicates of the data will exist in many places – spreadsheets, personal databases; the data warehouse is actually likely to reduce the unstructured duplication. For the end user a suite of tools is available to manipulate and present results from the data stored. These tools, commonly referred to as OLAP (on-line Analytical Processing) tools, are extremely proficient at drill-down, pivot, and presentation of results. They also have computational ability suitable for

things such as ratios and trends, although such functionality is simple compared to the calculations used in insurance for projections and reserving.

2.1 Data Mining Research Opportunities and Challenges

Data mining is one of the hottest topics in information technology. This section provides an introduction to data mining: what it is, why it is important, and how it can be used to provide increased understanding of critical relationships in rapidly expanding corporate data warehouse.

The field of data mining and knowledge discovery is emerging as a new, fundamental research area with important applications in science, engineering, medicine, business, and education. Data mining attempts to formulate, analyze, and implement basic induction processes that facilitate the extraction of meaningful information and knowledge from unstructured data. Data mining extracts patterns, changes, associations, and anomalies from large datasets. Work in data mining ranges from theoretical work on the principles of learning and mathematical representations of data to building advanced engineering systems that perform information filtering on the Web, find genes in DNA sequences, help understand trends and anomalies in economics and education, and detect network intrusion. Data mining is also a promising computational paradigm that enhances traditional approaches to discovery and increases the opportunities for breakthroughs in the understanding of complex physical and biological systems. Researchers from many intellectual communities have much to contribute to this field. These include the communities of machine, intellectual communities have much to contribute to this field. These include the communities of machine learning, statistics, databases, visualization and graphics, optimization, computational mathematics, and the theory of algorithms.

The amount of digital data has been exploding during the past decade, while the number of scientists, engineers, and analysts available to analyze the data has been static. To bridge this gap requires the solution of fundamentally new research problems, which can be grouped into the following broad challenges: A) developing algorithms and systems to mine large, massive, and high-dimensional datasets; B) developing algorithms and systems to mine new types of data; C) developing algorithms, protocols, and other infrastructure to mine distributed data; and D) improving the ease of use of data mining systems; E) developing appropriate privacy and security models for data mining.

There is an important need for support by government and business of basic, applied, multidisciplinary, and interdisciplinary research in data mining and knowledge discovery in order to respond to these challenges.

There is an important experimental component to data mining and knowledge discovery, which requires the creation and maintenance of appropriate systems, computational infrastructures, and test beds.

Data mining is the semiautomatic discovery of patterns, association, changes, anomalies, rules, and statistically significant structures and events in data. That is, data mining attempts to extract knowledge from data.

Data mining differs from traditional statistics in several ways: formal statistical inference is assumption driven in the sense that a hypothesis is formed and validated against the data. Data mining in contrast is discovery driven in the sense that patterns and hypothesis are automatically extracted from data. To put it in another way, data mining is data driven, while statistics is human driven. The branch of statistics that data mining resembles most is exploratory data analysis, although this field, like most of the rest of statistics, has focused on datasets far smaller than most that are the target of data mining researchers.

Data mining also differs from traditional statistics in that sometimes the goal is to extract qualitative models, which can easily be translated into logical rules or visual representations; in this sense data mining is human centered and is sometimes coupled with human–computer interface research.

Data mining is a step in the KDD process, which is an interactive, semiautomated process, which begins with raw data. Results of the data mining process may be insights, rules, or predictive models.

The field of data mining draws upon several subjects, including statistics, machine learning, databases, and high-performance computing.

In this section, we are primarily concerned with large datasets, massive datasets, and distributed datasets. By large, we mean datasets, which are too large to fit into the memory of a single workstation. By massive we mean datasets, which are too large to fit into the disks of a single workstation or a small cluster of workstations. Instead, massive clusters or tertiary storage such as tape are required. By distributed, we mean datasets that are geographically distributed.

The focus on large datasets is not just an engineering challenge; it is an essential feature of induction of expressive representations from raw data. It is only by analyzing large datasets that we can produce accurate logical descriptions, which can be translated automatically into powerful predictive mechanisms. Otherwise, statistical and machine learning principles suggest the need for substantial user input (specifying metaknowledge necessary to acquire highly predictive models from small datasets).

There are probably as many definitions of the term *data mining* as there are software analytical tool vendors in the market today. As with OLAP, which could mean almost anything, vendors and industry analysts have adopted the term *data mining* somewhat indiscriminately. The result is a blanket definition that includes all tools employed to help users analyze and understand their data. In this section, we explore a more narrow definition. Data mining is a set of techniques used in an automated approach to exhaustively explore and bring to the surface complex relationships in very large datasets. We discuss only datasets that are largely tabular in nature, having most likely been implemented in relational database management technology. However, these techniques can be, have been, and will be applied to other data

representations, including spatial data domains, text-based domains, and multimedia (image) domains.

A significant distinction between data mining and other analytical tools is in the approach they use in exploring the data inter-relationships. Many of the analytical tools available support a verification-based approach, in which the user hypothesizes about specific by relying on the intuition of the analyst to pose the original question and refine the analysis based on the results of potentially complex queries against a database. The effectiveness of this verification-based analysis is limited by a number of factors, including the ability of the analyst to pose appropriate questions and quickly return results, manage the complexity of the attribute space, and think "out of the box."

Most available analytical tools have been optimized to address some of these issues. Query-and-reporting tools address ease of use, letting users develop SQL queries through point-and-click interfaces. Statistical analysis packages provide the ability to explore relationships among a few variables and determine statistical significance against a population. Multidimensional and relational OLAP tools precompute hierarchies of aggregations along various dimensions in order to respond quickly to users' inquiries. New visualization tools let users explore higher dimensionality relationships by combining spatial and nonspatial attributes (location, size, color, and so on).

Data mining, in contrast to these analytical tools, uses discovery-based approaches in which pattern matching and other algorithms are employed to determine the key relationships in the data. Data mining algorithms can look at numerous multidimensional data relationships concurrently, highlighting those that are dominant or exceptional.

2.1.1 Recent Research Achievements

The opportunities today in data mining rest solidly on a variety of research achievements, the majority of which were the results of government-sponsored research. In this section, we study a few of the more important ones. Note that several of them are interdisciplinary in nature, resting on discoveries made by researchers from different disciplines working together collaboratively.

Neural Networks. Neural networks are systems inspired by the human brain. A basic example is provided by a back propagation network, which consists of input nodes, output nodes, and intermediate nodes called *hidden nodes*. Initially, the nodes are connected with random weights. During the training, a gradient descent algorithm is used to adjust the weights so that the output nodes correctly classify data presented to the input nodes. The algorithm was invented independently by several groups of researchers.

Tree-Based Classifiers. A tree is a convenient way to break large datasets into smaller ones. By presenting a learning set to the root and asking questions at each interior node, the data at the leaves can often be analyzed very simply. For example, a classifier to predict the likelihood that a credit card transaction is fraudulent may use an interior node to divide a training dataset into two

sets, depending upon whether or not five or fewer transactions were processed during the previous hour. After a series of such questions, each leaf can be labeled fraud/number-fraud by using a simple majority vote. Tree-based classifiers were independently invented in information theory, statistics, pattern recognition, and machine learning.

Graphical Models and Hierarchical Probabilistic Representations. A directed graph is a good means of organizing information on qualitative knowledge about conditional independence and causally gleaned from domain experts. Graphical models generalize Markov models and hidden Markov models, which have proved themselves to be a powerful modeling tool. Graphical models were independently invented by computational probabilists and artificial intelligence researchers studying uncertainty.

Ensemble Learning. Rather than use data mining to build a single predictive model, it is often better to build a collection or ensemble of models and combine them, say with a simple, efficient voting strategy. This simple idea has now been applied in a wide variety of contexts and applications. In some circumstances, this technique is known to reduce variance of the predictions and therefore to decrease the overall error of the model.

Linear Algebra. Scaling data mining algorithms often depends critically upon scaling underlying computations in linear algebra. Recent work in parallel algorithms for solving linear system and algorithms for solving sparse linear systems in high dimensions is important for a variety of data mining applications, ranging from text mining to detecting network intrusions.

Large-Scale Optimization. Some data mining algorithms can be expressed as large-scale, often nonconvex, optimization problems. Recent work has been providing parallel and distributed methods for large-scale continuous and discrete optimization problems, including heuristic search methods for problems too large to be solved exactly.

High-Performance Computing and Communication. Data mining requires statistically intensive operations on large datasets. These types of computations would not be practical without the emergence of powerful and high-performance clusters of workstations supporting protocols for high-performance computing such as MPI and MPIO. Distributed data mining can require moving large amounts of data between geographically separated sites, something that is now possible with the emergence of wide area high-performance networks.

Databases, Data Warehouse, and Digital Libraries. The most time-consuming part of the data mining process is preparing data for data mining. This step can be streamlined in part if the data is already in a database, data warehouse, or digital library, although mining data across different databases. Some algorithms, such as association algorithms, are closely connected to databases, while some of the primitive operations being built into tomorrow's data warehouses should prove useful for some data mining applications.

Visualization of Massive Datasets. Massive datasets often generated by complex simulation programs, required graphical visualization methods for best comprehension. Recent advances in multiscale visualization have allowed the rendering to be done far more quickly and in parallel, making these visualization tasks practical.

2.1.2 Data Mining Application Areas

Data mining techniques have been applied successfully in many areas from business to science and sports. Data mining has been used in database marketing, retail data analysis, stock selection, credit approval, etc. Data mining techniques have been used in astronomy, molecular biology, medicine, geology, and many more fields. It has also been used in health care management, tax fraud detection, money laundering monitoring, and even sports.

Market management. Target marketing, customer relationship management, market basket analysis, cross-selling, market segmentation.

Risk management. Forecasting, customer retention, improved underwriting, quality control, competitive analysis.

Fraud management. Fraud detection.

Industrial-specific applications. Banking, finance, and securities: Profitability analysis (for individual officer branch, product, product group, monitoring marketing programs and channels, customer data analysis customer segmentation profiling).

Telecommunications and media. Response scoring, marketing campaign management, profitability analysis, and customer segmentation.

Health care. FAMS (Fraud and Abuse Management System) assisting health insurance organizations dealing with fraud and abuse: detection, investigation, settlement, prevention of recurrence.

New Applications

The discipline of data mining is driven in part by new applications, which require new capabilities not currently being supplied by today's technology. These new applications can be naturally divided into three broad categories:

Business & E-commerce Data. Back office, front office, and network applications produce large amounts of data for business processes. Using this data for effective decision making remains a fundamental challenge.

Scientific, Engineering, and Health Care Data. Scientific data and metadata tend to be more complex in structure than business data. In addition, scientists and engineers are making increasing use of simulation and systems with application domain knowledge.

Web data. The data on the Web is growing not only in volume but also in complexity. Web data now includes not only text and image, but also streaming data and numerical data.

In this section, we describe several such applications from each category.

Business transactions. Today, businesses are consolidating and more and more businesses have millions of customers and billions of their transactions. They need to understand risks (Is this transaction fraudulent? Will the customers pay their bills?) and opportunities (What is the expected profit of this customers? What product is this customer most likely to buy next?).

Electronic commerce. Not only does electronic commerce produce large datasets in which the analysis of marketing patterns and risks patterns is critical, but unlike some of the applications above, it is also important to do this in real or near-real time, in order to meet the demand of on-line transactions.

Genomic data. Genomic sequence and mapping efforts have produced a number of databases, which are accessible over the Web. In addition, there is also a wide variety of other on-line databases, including those containing information about disease, cellular function, and drugs. Finding relationship between these data sources, which are largely unexplored, is another fundamental data mining challenge. Recently, scalable techniques have been developed for comparing whole genomes.

Sensor data. Satellites, buoys, balloons, and a variety of other sensors produce voluminous amounts of data about the earth's atmosphere, oceans, and lands. A fundamental challenge is to understand the relationships, including casual relationships amongst this data. For example, do industrial pollutants affect global warming? There are also large terabyte to petabyte datasets being produced by sensors and instruments in other disciplines, such as astronomy, high-energy physics, and nuclear physics.

Simulation Data. Simulation is now accepted as a third mode of science, supplementing theory and experiment. Today, not only do experiments produce huge datasets, but so do simulations. Data mining and more generally data intensive computing is proving to be a critical link between theory, simulation, and experiment.

Health care Data. Health care has been the most rapidly growing segment of the nation's gross domestic product (GDP) for some time. Hospitals, health care organizations, insurance companies, and the federal government have large collections of data about patients, their health care problems, the clinical procedures used, their costs, and the outcomes. Understanding relationships in this data is critical for a wide variety of problems, ranging from determining what procedures and clinical protocols are most effective to how best to deliver health care to most people in an era of diminishing resources.

Multimedia Documents. Few people are satisfied with today's technology for retrieving documents on the Web, yet the number of documents and the number of people accessing these documents is growing explosively. In addition, it is becoming increasingly easier to archive multimedia data, including audio, images, and video data, but progressively harder to extract meaningful information from the archives as the volume grows.

The Data Web. Today the Web is primarily oriented toward documents and their multimedia extensions. HTML has proved itself to be a simple, yet powerful language for supporting this. Tomorrow the potential exists for the Web to prove to be equally important for working with data. The extension markup language (XML) is an emerging language for working with data in networked environments. As this infrastructure grows, data mining is expected to be a critical enabling technology for the emerging data Web.

2.1.3 Success Stories

In this section, we briefly describe some success stories involving data mining and knowledge discovery.

Association Rules. Suppose we have a collection of items. The data for many applications consists of multiple of one or more items. A basic example is provided by a supermarket where the items are the products offered for sale and the transactions are purchases, consisting of one or more products purchased by an individual at a given time. A fundamental problem is to uncover associations: which products tend to be purchased together. There has been a lot of recent work on this problem and a variety of algorithms have been developed that can discover associations, even in very large datasets, with just a few passes over the data. A variety of commercial data mining systems support association rules and they are now routinely applied to a range of problems from database marketing to product placement for supermarkets. In addition, association rules algorithms have spurred new research in a variety of areas from databases to complexity theory.

Fraud Detection. Although relatively few credit card transactions are fraudulent, the sheer volume of transactions means that over $500 million are lost each year in this way. A variety of data mining techniques have been used to develop fraud systems, which can detect fraudulent credit card transactions in near-real time. This problem is challenging due to the size of the datasets, the rarity of the events of interest, and the performance requirements for near-real time detection. Data mining has also improved fraud detection in other application areas, including telecom fraud and insurance fraud.

Astronomical Data. Traditionally, the search of new galaxies, stars, and quasars has primarily been done by astronomers visually examining individual photographic plates. Classification algorithms from data mining have recently

been used to automate this process yielding new astronomical discoveries. The classification algorithms are applied to derive attributes produced by image processing, such as the brightness, area, and morphology of sky objects. The approach has also proved useful for detecting new objects too faint to be observed by a manual analysis or traditional computational techniques. For the 2nd Palomar Observatory Sky Survey, this approach resulted in over a three-fold increase in the size of the catalog.

Genomic Data. Genomic data is stored all over the world, in a variety of formats and managed by a variety of applications and systems. Recently, systems have been developed that allow discoveries to be made involving information distributed over several systems. In particular, the new systems have enabled for the first time whole genome comparison, gene identification, and whole genome functional interpretation and analysis. The techniques developed for analyzing genomic and other types of scientific data can be expected to play a role in analyzing a broad range of biological data.

Distributed Data Mining. Traditionally, data mining has required that the relevant data be warehoused in a single location. Recently, distributed data mining systems have been exploiting wide area, high-performance net networks, such as the NSF vBNS network, to mine large amounts of distributed scientific and health care data. Recently, these systems have been setting records for the sustained movement of very large amounts of data over wide area networks. Separately, a prototype has been developed by exploiting distributed data mining to improve the detection of credit card fraud.

Text Mining. Recently, data mining has been combined with algorithms from information retrieval to improve the precision and recall the queries on very large collections of documents. In particular, some of these algorithms have proved useful in multilingual collections and others have shown their worth in querying using concepts instead of keywords.

2.1.4 Trends that Affect Data Mining

In this section, we describe five external trends, which promise to have a fundamental impact on data mining.

Data Trends . Perhaps the most fundamental external trend is the explosion of digital data mining during the past two decades. During this period, the amount of data probably has grown between six and ten orders of magnitude. Much of this data is accessible via networks. On the other hand, during the same period the number of scientists, engineers, and other analysts available to analyze this data has remained relatively constant. For example, the number of Ph.D.s in statistics graduating each year has remained relatively constant during this period. Only one conclusion is possible: either most of the data is destined to be write-only; or techniques, such as data mining, must be developed, which can automate, in part, the analysis of this data, filter irrelevant information, and extract meaningful knowledge.

Hardware Trends. Data mining requires numerically and statistically intensive computations on large datasets. The increasing memory and processing speed of workstations enables the mining of dataset using current algorithms and techniques that were too large to be mined just a few years ago. In addition, the commoditization of high-performance computing through SMP workstations and high-performance workstation clusters enables attacking data mining problems that were accessible using only the largest supercomputers of few years ago.

Network Trends. The next generation Internet (NGI) will connect sites at OC-3 (155 MBits/sec), speeds and higher. This is over 100 times faster than the connectivity provided by current networks. With this type of connectivity, it becomes possible to correlate distributed datasets using current algorithms and techniques. In addition, new protocol, algorithms, and languages are being developed to facilitate distributed data mining using current and next generation networks.

Scientific Computing Trends. As mentioned above, scientists and engineers today view simulation as a third mode of science. Data mining and knowledge discovery serve an important role linking the three modes of science: theory, experiment, and simulation, especially for those cases in which the experiment or simulation results in large datasets.

Business Trends. Today businesses must be more profitable, react quicker, and offer higher quality services than ever before, and do it all using fewer people and at lower cost. With these types of expectations and constraints, data mining becomes a fundamental technology, enabling business to more accurately predict opportunities and risks generated by their customers and their customer's transactions.

2.1.5 Research Challenges

In this section, we describe some of the major research challenges identified by the three workshops. The research challenges are divided into five broad areas: A) improving the scalability of data mining algorithms, B) mining nonvector data, C) mining distributed data, D) improving the ease of use of the data mining systems and environments, and E) privacy and security issues for data mining.

A. *Scaling data mining algorithms.* Most data mining algorithms today assume that the data fits into memory. Although success on large datasets is often claimed, usually this is the result of sampling large datasets until they fit into memory. A fundamental challenge is to scale data mining algorithms as
the number of records or observations increases;
the number of attributes per observation increases,

the number of predictive models or rule sets used to analyze a collection of observation increases; and,
as the demand for interactivity and real-time response increases.
Not only must distributed, parallel, and out-of-memory versions of current data mining algorithms be developed, but genuinely new algorithms are also required. For example, association algorithms today can analyze out-of-memory data with one or two passes, while requiring only some auxiliary data be kept in memory.

B. *Extending data mining algorithms to new types.* Today, most data mining algorithms work with vector-valued data. It is an important challenge to extend data mining algorithms to work with other data types, including 1) time series and process data, 2) unstructured data, such as text, 3) semistructured data, such as HTML and XML documents, 4) multimedia and collaborative data, 5) hierarchical and multiscale data, and 6) and collection-valued data.

C. *Developing distributed data mining algorithms.* Today most data mining algorithms require bringing all together data to be mined in a single, centralized data warehouse. A fundamental challenge is to develop distributed versions of data mining algorithms, so that data mining can be done while leaving some of the data in place. In addition, appropriate protocols, languages, and network services are required for mining distributed data to handle the metadata and mappings required for mining distributed data. As wireless and pervasive computing environments become more common, algorithms and systems for mining the data produced by these types of systems must also be developed.

D. *Ease of use.* Data mining today is at best a semiautomated process and perhaps destined to always remain so. On the other hand, a fundamental challenge is to develop data mining systems, which are easier to use, even by casual users. Relevant techniques include improving use interface, supporting casual browsing and visualization of massive and distributed datasets, developing techniques and systems to manage the metadata required for data mining, and developing appropriate languages and protocols for providing causal access to data. In addition, the development of data mining and knowledge discovery environments that address the process of collecting, processing, mining, and visualizing data, as well as the collaborative and reporting aspects necessary when working with data and information derived from it, is another important fundamental challenge.

E. *Privacy and Security.* Data mining can be a powerful means of extracting useful information from data. As more and more digital data becomes available, the potential for misuse of data mining grows. A fundamental challenge is to develop privacy and security models and protocols appropriate for data mining and to ensure that next generation data mining systems and designed from the ground up to employ these models and protocols.

2.1.6 Test Beds and Infrastructure

Experimental studies will play a critical role in advancing the field of data mining. Developed test beds for high performance and distributed data mining is essential for progress in the field.

The requirements for data mining test beds are different than those for general purpose high-performance computing test beds. For example, the computing resources for data mining test beds are as much disk oriented as processor oriented; the network resources must be able to move datasets and data elements between geographically distributed sites with guaranteed quality of service, and a variety of general purpose and specialized data mining software must be available.

Perhaps the two most difficult challenges in creating data mining test beds and national resources in data mining are assembling a) the appropriate datasets and b) the required interdisciplinary and multidisciplinary teams.

2.1.7 Findings and Recommendations

In this section, we list some of the major findings and recommendations.For all Interested Parties

Data mining and knowledge discovery is a new emerging discipline with both a scientific and an engineering component that is of strategic importance for the U.S.A. and of critical importance to future information access technologies. All interested parties are encouraged to work toward the maturation of data mining and knowledge discovery, its establishment as a scientific and engineering discipline in its own right, and the evolution of a community that includes the relevant traditions and disciplines and put them together in the proper context.

For the Federal Government

Create programs that encourage the emergence of data mining and knowledge discovery as an independent discipline. Support interdisciplinary and multidisciplinary research projects. Many advances in data mining require teams of mathematical and statisticians, computer scientists, and application domain scientists working together to create the appropriate datasets and the required algorithms and software to analyze them.

Support basic research in computer and information sciences that underlies mining, including machine learning, knowledge systems, databases' high-performance computing, high-performance networking, and digital libraries.

Support basic research in mathematics and statistics that underlies data mining, including statistics, probability, applied mathematics, logic, discrete mathematics, analysis and dynamical systems, linear algebra, and computational geometry and algebra.

Support data mining test beds:

The hardware, software, data and consulting requirements for data mining often outstrip the resources of individual scientists and small research groups. Supporting national resources and test beds for data mining is important in order to provide the proper experimental infrastructure required for next generation data mining experiments.

For Companies

Support applied research in data mining:

Work to develop, implement, and support appropriate privacy and security models for data mining systems.

Create sanitized versions of real datasets for use by data mining researchers.

Support joint research projects between industry and universities. Support collaborative test beds and demonstration projects.

For Scientists and Engineers:

As new data is collected and archived, support emerging protocols, languages, and standards to facilitate the future analysis and mining of the data, especially by scientists and engineers from other disciplines.

As new data is collected and new systems are built to manage it, ensure that the best available privacy and security models are used to protect inadvertent disclosures of private information.

Provide long-term maintenance and access to datasets created by scientists and engineers, as well as to the knowledge and information extracted from them.

Data mining and knowledge discovery are emerging as a new discipline with important applications to science, engineering, health care, education, and business. Data mining rests firmly on 1) research advances obtained during the past two decades in a variety of areas and 2) more recent technological advances in computing, networking, and sensors. Data mining is driven by the explosion of digital data and the scarcity of scientists, engineers, and domain experts available to analyze it.

Data mining is beginning to contribute research advances of its own, by providing scalable extension and advances to work in associations, ensemble learning, graphical models, techniques for on-line discovery, and algorithms for the exploration of massive and distributed datasets.

Advances in data mining require supporting a) single investigators working in data mining and the underlying research domains supporting data mining; b) interdisciplinary and multidisciplinary research groups working on important basic and applied data mining problems; and c) the appropriate test beds for mining large, massive, and distributed datasets. Appropriate privacy and security models for data mining must be developed and implemented.

2.2 Evolving Data Mining into Solutions for Insights

The capacity of digital data storage worldwide has doubled every nine months for at least a decade, at twice the rate predicted by Moore's law for the growth of computing power during the same period. This less familiar but noteworthy phenomenon, which we call storage law, is among the reasons for the increasing importance and rapid growth of the field of data mining.

The aggressive rate growth of disk storage and the gap between Moore's law and storage law growth trends represents a very interesting pattern in the state of technology evolution. The ability to capture and store data has far outpaced our ability to process and utilize it. This growing challenge has produced a phenomenon we call *data tombs*, or *data stores* that are effectively write-only; data is deposited to nearly rest in peace, since in all likelihood it will never be accessed again.

Data tombs also represent missed opportunities where the data might support exploration in a scientific activity or commercial exploitation by a business organization, the data is potentially valuable information. Without next generation data mining tools, most will stay unused; hence most of the opportunity to discover, profit, improve service, or optimize operations will be lost. Data mining – one of the most general approaches in reducing data in order to explore analyze and understood it – is the focus of this special section.

Data mining is defined as the identification of interesting structure in data. Structure designates patterns, statistical or predictive modes of the data, and relationships among part of the data. Each of these terms – patterns, models, and relationships – has a concrete definition in the context of data mining. A pattern is a parsimonious summary of a subset of the data (such as people who own minivans have children). A model of the data can be a model of the entire dataset and can be predictive; it can be used to, say, anticipate future customer behavior (such as the likelihood a customer is or is not happy, based on historical data of interaction with a particular company). It can also be a general model (such as a joint probability distribution or set of variables in the data). However, the concept of interesting is much more difficult to define.

What structure within a particular dataset is likely to be interesting to a user or task? An algorithm could easily enumerate lots of patterns from a finite database. Identifying interesting structure and useful patterns among the plethora of possibilities is what a data mining algorithm must do, and it must do it quickly over very large databases.

For example, frequent item sets (variable values occurring together frequently in a database of transactions) could be used to answer, say, which items are most frequently bought together in the same supermarket. Such an algorithm could also discover a pattern in a demographics database with exceptionally high confidence that, say, all husbands are males. While true, however, this particular association is unlikely to be interesting. This same method did uncover in the set of transactions representing physicians billing the Australian Government's medical insurance agency a correlation deemed

extremely interesting by the agency auditors. Two billing codes were highly correlated: they were representative of the same medical procedure and hence had created the potential for double billing fraud. This nugget of information represented millions of dollars overpayment.

The quest for patterns in data has been studied for a long time in many fields, including statistics, pattern recognitions, and exploratory data analysis. Data mining is primarily concerned with making it easy, convenient, and practical to explore very large databases for organizations and users with lots of data but without years of training as data analysts. The goals uniquely addressed by data mining fall into certain categories:

Scaling analysis to large databases. What can be done to large datasets that cannot be loaded and manipulated in main memory? Can abstract data access primitives embedded in database systems provide mining algorithms with the information to drive a search for patterns? How might we avoid having to scan an entire very large database while reliably searching for patterns?

Scaling to high-dimensional data and models. Classical statistical data analysis relies on humans to formulate a model, and then use the data to access the models fit to the data. But humans are ineffective in formulating hypothesis when datasets have a large number of variables (possibly thousands in cases involving demographics and hundreds of thousands in cases involving retail transactions, Web browsing, and text document analysis). The model derived from this automated discovery and search process can be used to find lower dimensional subspaces where people find it easier to understand the aspects of the problem that are interesting.

Automating search. Instead of relying solely on human analysts to enumerate and create hypothesis, the algorithms perform much of this tedious and data-intensive work automatically.

Finding patterns and models understandable and interesting to users. Classical methodologies for scoring models focus on notions of accuracy (how well the model predicts the data) and utility (how to measure the benefit of the derived pattern, such as money saved). While these methods are well understood in decision analysis, the data mining community is also concerned with new measures, such as the understanding of a model or the novelty of a pattern and how to simplify a mode for interpretability. It is particularly important that the algorithms help end users gain insight from data by focusing on the extraction of patterns that are easily understood or can be turned into meaningful reports and summaries by trading off complexity for understandability.

2.2.1 Trends and Challenges

Among the most important trends in data mining is the rise of "verticalized," or highly specialized, solutions, rather than the earlier emphasis on building new data mining tools. Web analysis, customer behavior analysis, and customer relationship management all reflect the new trend; solutions to business

problems increasingly embed data mining technology, often in a hidden fashion, into the application. Hence, data mining applications are increasingly targeted and designed specifically for end users. This is an important and positive departure from most of the field's earlier work, which is used to focus on building mining tools for data mining experts.

Transparency and data fusion represent two important challenges for the growth of the data mining market and technology development. Transparency concerns the need for an end user-friendly interface, whereby the data mining is transparent as far as the user is concerned. Embedding vertical applications is a positive step toward addressing this problem, since it is easier to generate explanations from models built in a specific context. Data fusion concerns a more pervasive infrastructure problem: where is the data that has to be mined? Unfortunately, most efforts at building the decision-support infrastructure, including data warehouses, have proved to be big, complicated, and expensive. Industry analysts report failure of a majority of enterprise data warehousing efforts. Hence, even though the data accumulates in stores, it is not being organized in a format that is easy to access for mining or even for general decision support.

Much of the problem involves data fusion. How can a data miner consistently reconcile a variety of data sources? Often labeled as data integration, warehousing, or IT initiatives, the problem is also the unsolved prerequisite to data mining. The problem of building and maintaining useful data warehouses remains one of the great obstacles to successful data mining. The sad reality today is that before users get around to applying a mining algorithm, they must spend months or years bringing together the data sources. Fortunately, new disciplined approaches to data warehousing and mining are emerging as part of the vertical solutions approach.

2.3 Knowledge Extraction Through Data Mining

Data mining is the process of sifting through and analyzing rich sets of domain specific data and then extracting the information and knowledge in the form of new relationships, patterns, or clusters for decision making purposes. Thus data mining is a form of knowledge discovery essential for solving problems in a specific domain.

The term *KDD* denotes the overall process of extracting the high-level knowledge from low-level data. The multitude of terms used for KDD includes data or information harvesting, data archeology, functional dependency analysis, knowledge extraction, and data pattern analysis.

Conventionally *data mining* refers to the act of extracting patterns or models from data (be it automated or human assisted). However many steps precede the data mining step: retrieving the data from large warehouse (or some other source), selecting the appropriate subset to work with, deciding on the appropriate sampling strategy, cleaning the data and dealing with

missing fields, and applying the appropriate transformations, dimensionality reduction, and projections. The data mining step then fits models to or extracts patterns from, the preprocessed data. However, to decide whether this extracted information does represent knowledge, one needs to evaluate this information, perhaps visualize it, and finally consolidate it with existing (and possibly contradictory) knowledge. Obviously these steps are all on the critical path from data to knowledge. Furthermore any one step can result in changes in preceding or succeeding steps often requiring starting from scratch with new choices and settings. Hence in the definition, data mining is just a step in the overall KDD process.

The concept was born about ten years ago. The interest in the data mining field and its exploitation in different domains (marketing, finance, banking, engineering, health care, power systems, meteorology, etc.) has been increasingly recently due to a combination of factors. They include:

- the emergence of very large amount of data (terabytes – bytes – of data) due to computer-automated data measurement and/or collection, digital recording, centralized data archives, and software and hardware simulations.
- the dramatic cost decrease of mass storage devices
- the emergence and growth of fielded database management systems
- the advances in computer technology such as faster computers and parallel architectures
- the continuous developments in automatic learning techniques
- the possible presence of uncertainty in data (noise, outliers, missing information).

The general purpose of data mining is to process the information from the enormous stock of data we have or that we may generate, so as to develop better ways to handle data and support future decision making. Sometimes, the pattern to be searched for and the models to be extracted from data are subtle, and require complex calculus and/or significant specific domain knowledge. Or even worse, there are situations where one would like to search for patterns that humans are not well suited to find, even if they are good experts in the field. For example, in many power system-related problems one is faced with high-dimensional datasets that cannot be easily modeled and controlled on the whole, and therefore automatic methods capable of synthesizing structures from such data become a necessity.

This section presents the concept of data mining and aims at providing an understanding of the overall process and tools involved: how the process turns out, what can be done with it, what are the main techniques behind it, and which are the operational aspects. We aim also at describing a few examples of data mining applications, so as to motivate the power system field as a very opportune data mining application.

2.3.1 Data Mining Process

Data mining process consists of three major steps. Of course, it all starts with a big pile of data. The first processing step is data preparation often referred to as "scrubbing the data." Data is selected, cleaned, and preprocessed under the guidance and knowledge of a domain expert. The most time-consuming part of the data mining process is preparing data for data mining. This step can be streamlined in part if the data is already in a database, data warehouse, or digital library, although mining data across different databases, for example, is still a challenge. Second, a data mining algorithm is used to process the prepared data, compressing and transforming it to make it easy to identify any latent valuable nuggets of information. In the second step in data mining, once the data is collected and preprocessed, the data mining algorithms perform the actual sifting process. Many techniques have been used to perform the common data mining activities of associations, clustering, classification, modeling, sequential patterns, and time series forecasting. These techniques range from statistics to rough sets to neural networks.

The third phase is the data analysis phase, where the data mining output is evaluated to see if additional domain knowledge was discovered and to determine the relative importance of the facts generated by the mining algorithms.

The final step is the analysis of the data mining results or output. In some cases the output is in a form that makes it very easy to discern the valuable nuggets of information from the trivial or uninteresting facts. The relationships are represented in if-then rules form. With rules recast into textual form, the valuable information is much easier to identify. In other cases, however, the results will have to be analyzed either visually or through another level of tools to classify the nuggets according to the predicted value.

Whatever be the data mining algorithm used, the results will have to be presented to the user. A successful data mining application involves the transformation of raw data into a form that is more compact and more understandable, and where relationships are explicitly defined.

For most data mining applications, the relatively clean data that resides in the corporate data warehouse must usually be refined and processed before it undergoes the data mining process. This preprocessing might involve joining information from multiple tables, selecting specific rows or records of data, and it most certainly include selecting the columns or fields of data that need to be looked at in the data mining step. Often two or more fields are combined to represent ratios or derived values. This data selection and manipulation process is usually performed by someone with a good deal of knowledge about the problem domain and the data related to the problem under study. Depending on the data mining algorithm involved, that data might need to be formatted in specific ways (such as scaling of numeric

data) before it is processed. Hence data preparation is crucial to successful data mining application.

The data mining process consists of three major steps.

(1) Data Preparation: Data is selected, cleaned, and preprocessed under the guidance and knowledge of domain experts who capture and integrate both the internal and external data into a comprehensive view that encompasses the whole organization.
(2) Data mining algorithm: Data mining algorithm is used to mine the integrated data to enable easy identification of any valuable information.
(3) Data Analysis Phase: Data mining output is evaluated to see if the domain knowledge discovered is in the form of rules extracted out of the network. The overall data mining process is shown in Fig 2.1.

In general, once the data is prepared, a neural network is used to build a model based on the data. When the model is ready then it opens a way to complete automation of the process as shown in Fig 2.2.

Data mining consists of five major elements:

Extract, transform, and load transaction data onto the data warehouse system.
Store and manage the data in a multidimensional database system.
Provide data access to business analysts and information technology professionals.
Analyze the data by application software.
Present the data in a useful format, such as a graph or table.
The typical KDD process is illustrated in Fig. 2.3.

By definition, data mining is the nontrivial process of extracting valid, previously unknown, comprehensible, and useful information from large databases and using it. It is an exploratory data analysis, trying to discover useful patterns in data that are not obvious to the data user.

What is a database (DB)? It is a collection of objects (called tuples in the DB jargon, examples in machine learning or transactions in some application fields), each one of which is described by a certain number of attributes, which provide detailed information about each object. Certain attributes are selected as input attributes for a problem, certain ones as outputs (i.e., the desired objective: a class, a continuous value, etc.). Table 2.1 shows some examples

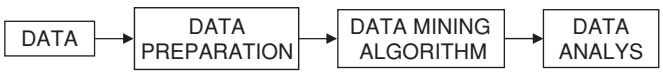

Fig. 2.1. Overall data mining process

Fig. 2.2. Automation of data mining process

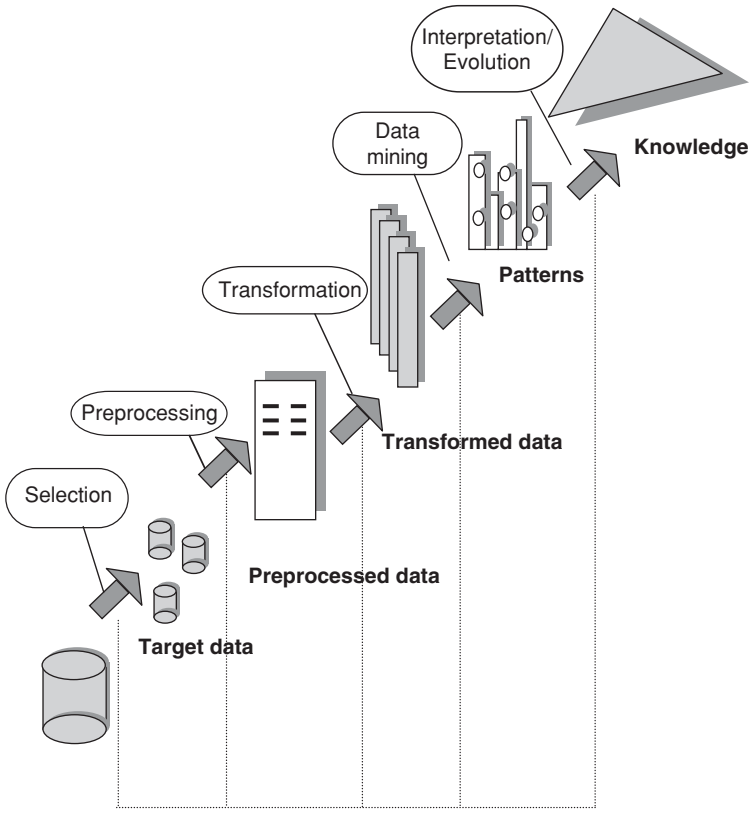

Fig. 2.3. A typical knowledge discovery process

of hourly energy transactions recorded in a data for a power market analysis application (each row of the table corresponds to an object and each column indicates one attribute, e.g., buyer, quantity, price). In such an application, the power system is considered as a price-based market with bilateral contracts (i.e., direct contracts between the power producers and users or broker outside of a centralized power pool), where the two parties, the buyer and the seller, could be utility distribution companies, utility and nonutility retailers (i.e., energy service providers), independent generators (i.e., independent power producers), generation companies, or end customers (as single customers or as parts of aggregated loads). The example will be used further in order to exemplify the techniques implied in the data mining.

Usually, one of the first tasks of a data mining process consists of summarizing the information stored in the database, in order to understand well its content. This is done by means of statistical analysis or query-and-reporting techniques. Then more complex operations are involved, such as identifying models, which may be used to predict information about future objects. The term *supervised learning* (known as "learning with a teacher") is implied in

Table 2.1. Example of a database

Buyer	Seller	Date	Hour ending	Product/ Service	Quantity	Unitary Price (%)	Transaction Price (Price Units)
A	B	23 Feb. 1998	9 a.m.	Energy	20 MWh	100	2000
A	C	23 Feb. 1998	11 a.m.	Energy	50 MWh	80	4000
D	A	5 Apr. 1998	9 a.m.	Energy	30 MWh	150	4500
A	B	9 Apr. 1998	2 p.m	Spinning Reserve	10 MW	100	1000
E	B	15 May 1998	4 a.m	Energy	30 MWh	70	2100
E	C	15 May 1998	5 a.m	Spinning Reserve	20 MW	200	4000
E	B	31 July 1998	8 a.m	Spinning Reserve	10 MW	100	1000

mining data in which for each input of the learning objects, the desired output objective is known and implicated in leaning. In unsupervised learning approaches ("learning by observation") the output is not provided or not considered at all, and the method learns by itself only from input attribute values.

Notice that generally only about 10% of the total collected data is ever analyzed (not only means of data mining). Many companies realize the poor quality of their data collection only when a data mining analysis is started on it. The databases are usually very expensive to create and maintain, and for a small additional investment in mining them, highly profitable information may be discovered hidden in the data. Thus, the classical scenario is as follows: a company realizing that there might be "nuggets" of information in the data they process starts building a long-term repository (a data warehouse) to store as much data as possible (e.g., by recording systematically all purchases by individual customers of a supermarket); then they would launch a pilot DM study in order to identify actual opportunities; finally some of the applications identified as interesting would be selected for actual implementation.

However, apart from the "cheaply" collected or already available data, there are some applications of data mining where the data is produced by computer simulations or expensive real experiments. For example, in the case where future yet unknown situations have to be forecast, or in fields where security aspects are analyzed for a system (computer system, power system, or banking system) when the history does not provide fortunately negative examples, one may use Monte Carlo simulations in order to generate a DB automatically and this is itself a nontrivial task.

The usual film when a company or a holder of a big amount of data decides that the information he has collected is worthy of being analyzed unfolds like

this: he comes with the data to the data miner (e.g., a consultant), the data miner first gets familiar with the field of application and with the application specifics, then depending on the data mining software he has, he will select a portion of the available data and apply those techniques he expects to give him more knowledge in terms of some established objectives. In case the results of this combination of tools do not give to the interested one any improvement in the existing knowledge about the subject, either the miner gives up (it is indeed possible that this process yields only uninteresting results), or he tries to go further by implementing new methods for mining the specific data (e.g., for a temporal problem of early anomalies detection, a temporal decision tree may offer more valuable results than a decision tree).

Who is a data miner? – some person, usually with background in computer science or in statistics and in the domain of interest, or a couple of specialists, one in data mining, one in the domain of interest, who is able to perform the steps of the data mining process. The miner is able to decide how much iterative the whole process is and to interpret the visual information he gets at every substep.

In general the data mining process iterates through five basic steps:

- *Data selection.* This step consists of choosing the goal and the tools of the data mining process, identifying the data to be mined, then choosing appropriate input attributes and output information to represent the task.
- *Data transformation.* Transformation operations include organizing data in desired ways, converting one type of data to another (e.g., from symbolic to numerical), defining new attributes, reducing the dimensionality of the data, removing noise, "outliers," normalizing, if appropriate, deciding strategies for handling missing data.
- *Data mining step per se.* The transformed data is subsequently mined, using one or more techniques to extract patterns of interest. The user can significantly aid the data mining method by correctly performing the proceeding steps.
- *Result interpretation and validation.* For understanding the meaning of the synthesized knowledge and its range of validity, the data mining application tests its robustness, using established estimation methods and unseen data from the database. The extracted information is also assessed (more subjectively) by comparing it with prior expertise in the application domain.
- *Incorporation of the discovered knowledge.* This consists of presenting the results to the decision maker who may check/resolve potential conflicts with previously believed or extracted knowledge and apply the new discovered patterns.

Figure 2.3 presents schematically the whole process, by showing what happens with the data: it is preprocessed, mined, and postprocessed, the result being a refinement in the knowledge about the application. The data mining process is iterative, interactive, and very much a trial-and-error activity.

Visualization plays an important role. Because we find it difficult to emulate human intuition and decision making on a machine, the idea is to transform the derived knowledge into a format that is easy for humans to digest, such as images or graphs. Then, we rely on the speed and capability of the human user visual system to spot what is interesting at every step of the data mining process: preliminary representation of data, domain specific visualization, or result presentation.

From the point of view of software structure, there are two types of possible implementation:

- The one called data mining "in place": the learning system accesses the data through a database management system (DBMS) and the user is able to interact with both the database (by means of queries) and the data mining tools. The advantage is that the approach may handle very large database and may exploit the DBMS (e.g., the handling of distributed data).
- The one called data mining "offline": the objects are first loaded in the data mining software, with a translation into a particular form, outside the database, and the user interacts mainly with the data mining software. They allow to use the existing machine learning systems with only minor modification in implementation, and it may be faster but are generally limited to handle medium-sized datasets, which can be represented in main memory (up to several hundred Mbytes)

What can be done in the Data Mining step? - Depending mainly on the application domain and the interest of the miner, one can identify several types of data mining tasks for which data mining offers possible answers. We present them so that they are usually implied in the process. Possible results for each one of these tasks are provided by considering the example in table 1 as the database to be mined:

Summarization: Summarization is the abstraction or generalization of the data. A set of task relevant data is summarized and abstracted. This results in a smaller set, which gives a general overview of the data, usually with aggregate information. It aims at producing compact and characteristic description for a given set of data. It can take multiple forms: numerical (simple descriptive statistical measures like means, standard deviations, and so on), graphical (histograms, scatter plots, to name a few), or the form of "if-then" rules. It may provide descriptions about objects in the whole database or in selected subsets. *Example of summarization*: *"the minimum unitary price for all the transactions with energy is 70 price units" (see table.1).*

Clustering. A clustering problem is an unsupervised learning problem, which aims at finding in the data clusters of similar objects sharing a number of interesting properties. It may be used in data mining to evaluate similarities among data, build a set of representative prototypes, analyze correlations between attributes, or automatically represent a dataset by a small number of regions, preserving the topological properties of the original input space.

Example of a clustering result: "from the seller B point of view, buyers A and E are similar customers in terms of total price of the transactions done in 1998".

Clustering approaches address segmentation problems. These approaches assign records with a large number of attributes into a relatively small set of groups or "segments." This assignment process is performed automatically by clustering algorithms that identify the distinguishing characteristics of the dataset and then partition the n-dimensional space defined by the dataset attributes along natural leaving boundaries. There is no need to identify the groupings desired or the attributes that should be used to segment the dataset.

Clustering is often one of the first steps in the data mining analysis. It identifies groups of related records that can be used as a starting point for exploring further relationships. This technique supports the development of population segmentation models, such as demographic-based customer segmentation. Additional analyses using standard analytical and other data mining techniques can determine the characteristics of these segments with respect to some desired outcome. For example, the buying habits of multiple population segments might be compared to determine those segments that need to be targeted for a new sales campaign.

Classification. A classification problem is a supervised learning problem where the output information is a discrete classification, i.e., given an object and its input attributes the classification output is one of the possible mutually exclusive classes of the problem. The aim of the classification task is to discover some kind of relationship between the input attributes and the output class, so that the discovered knowledge can be used to predict the class of a new unknown object. *Example of a derived knowledge rule, which classifies sales made early in the day (a sale is said to be early if it was made between 6 a.m. and 12 a.m.): "if the product is energy then the sale is likely to be early (confidence 0.75)".*

This technique is used to classify database records into a number of predefined classes based on certain criteria. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. Classification involves associating an observation with one of several labels called classes. Classification provides a mapping from attributes to specified groupings. Once the data is classified the traits of these specific groups can be summarized. In this chapter also, pattern classification is done to identify the specific class of each test pattern.

Classification, perhaps the most commonly applied data mining technique, employs a set of preclassified examples to develop a model that can classify the population of records at large. Fraud detection and credit-risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The use of classification algorithms begins with a training set of preclassified example transactions. For a fraud detection application, this would include complete records of both fraudulent and valid activities, determined on a

record-by-record basis. The classifier training algorithm uses these preclassified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called *classifier*.

The approach affects the explanation capability of the system. Once an effective classifier is developed, it is used in a predictive mode to classify new records into these same predefined classes. For example, a classifier capable of identifying risky loans could be used to aid in the decision of whether to grant a loan to an individual.

Regression. A regression problem is a supervised learning problem of building a more or less transparent model, where the output information is a continuous numerical value or a vector of such values rather than a discrete class. Then given an object, it is possible to predict one of its attributes by means of the attributes, by using the built model. The prediction of numerical values may be done by classical or more advanced statistical methods and by "symbolic" methods often used in the classification task. *Example of a model derived in a regression problem: "when buyer A buys energy, there exists a linear dependence between the established unitary price and the quantity he buys."*

Dependency modeling. A dependence modeling problem consists in discovering a model, which describes significant dependencies among attributes. These dependencies are usually expressed as "if-then" rules in the form "if antecedent is true then consequent is true," where both the antecedent and the consequent of the rule may be any combination of attributes, rather than having the same output in the consequent like in the case of the classification rules. *Example: such a rule might be "if product is energy then transaction price is larger than 2000 price units."*

Deviation detection. This is the task focusing in discovering the most significant changes or deviations in the data between the actual content of the data and its expected content (previously measured) or normative values. It includes searching for temporal deviations (important changes in data with time) and group deviations (unexpected differences between two subsets of data). In our example, deviation detections could be used in order to find main differences between sales patterns in different periods of the year.

Temporal problems. In certain applications it is useful to produce rules, which take into account explicitly the role of time. There are databases containing temporal information, which may be exploited by searching for similar temporal patterns in data or learn to anticipate some abnormal situations in data. *Example*: *"a customer buying energy will buy spinning reserve later on (confidence 0.66)," or "if total quantity of daily transactions is less than 100 piece units during at least 1 month for a client, the client is likely to be lost."*

Causation modeling. This is a problem of discovering relationships of cause and effect among attributes. A casual rule of type "if-then" indicates not only that there is a correlation between the antecedent and the consequent of the

rule, but also that the antecedent causes the consequent. *Example: "decreasing energy price will result in more sold energy daily."*

What techniques are behind all these tasks? - The enumerated types of data mining tasks are based on a set of important techniques originating in artificial intelligence paradigms, statistics, information theory, machine learning, reasoning with uncertainty (fuzzy sets), pattern recognition, or visualization. Thus, a data mining software package is supported to varying degrees by a set of technologies, which nearly always includes:

- *Tree and rule induction.* Machine learning (ML) is the center of the data mining concept, due to its capabilities to gain physical insight into a problem, and participate directly in data selection and model search steps. To address problems like classifications (crisp and fuzzy decision trees), regression (regression tree), time-dependent prediction (temporal trees), ML field is basically concerned with the automatic design of "if-then" rules similar to those used by human experts. Decision tree induction, the best known ML framework, was found to be able to handle large scale problems due to its computational efficiency, provide interpretable results and in particular, able to identify the most representative attributes for a given task.
- *Association rules.* These techniques identify affinities among the collection as reflected in the examined records. These affinities are often expressed as rules. Given a set of transactions, where each transaction is a set of literal, an association rule is a set of literals, the association rule is an expression of the form X $\Rightarrow$ Y, where X and Y are set of items. The intuitive meaning of such a rule is that a transaction of the database that contains X tends to contain Y. Association rule generators are a powerful data mining technique used to search through an entire dataset, for rules revealing the nature and frequency of relationships or associations between data entities. The resulting associations can be used to filter the information for human analysis and possibly to define a prediction model based on observed behavior.
- *Clustering methods.* They are used often in the data selection preprocessing step, due to the property of learning unsupervised similarities between objects and reducing the search space to a set of most important attributes for the application, or to a finite set of objects "alike." The most frequently used clustering method is the *k-means* method, which identifies a certain number of groups of similar objects; it may be used in combination with the *nearest-neighbor rule*, which classifies any new object in the group most similar (most near) to it. This method may also be used in order to identify outliers in a database. For example, by using this technique it might be possible in our example to identify groups of similar sales (large quantity and cheap unitary price versus small quantity and expensive unitary price) and to find out that some of the sales are outliers (e.g., small quantity and cheap). Then a supervised learning technique might be used in order to

find a rule to characterize these abnormal sales, in terms of attributes (seller, buyer, product, date, etc.).

- *Artificial neural networks.* Nonlinear predictive models learn through training and resemble biological neural networks in structure. Neural network has probably been of greater interest through the formative stages of data mining technology. True neural networks are biological systems that detect patterns, make predictions, and learn. Artificial neural networks are computer programs implementing sophisticated pattern detection. Although artificial neural networks cannot completely mimic the human brain and have some limitations, they have the advantage of being a highly accurate predictive model, which can be applied across a large number of problems.

They are recognized in the automatic learning framework as "universal approximators," with massively parallel computing character and good generalization capabilities, and also as black boxes due to the difficulty in obtaining insight into the relationship learned. They are used within the data mining step: to generate a regression model that can predict future behavior, on the basis of a database with input–output pairs of continuous numerical historical information (the neural network acts like a mapping, associating numerical outputs to any new object of known attributes values), and to automatically represent a dataset by a small number of representative prototypes, preserving the topological properties of the original attribute space (unsupervised learning).

- *Statistical techniques such as linear regression*, discriminant analysis, or statistical summarization. Classical statistical approaches include Bayesian network, regression analysis, correlation analysis, and cluster analysis. Modern statistical approach is the nearest neighbor method. An optimal model, based on a defined statistical measure, is searched among the patterns and regularities are then drawn from the model. K-nearest neighbor technique classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where $k > 1$). It is sometimes called the *k-nearest neighbor* method.
- *Visualization techniques*: In this method data is transformed into visual objects such as dots, lines, and areas. The data is then displayed in a two- or three-dimensional space. Users can interactively explore the interesting spots by visual examination. In the visual interpretation of complex relationships in multidimensional data, graphic tools are used to illustrate data relationships.

Visualization of Massive Datasets: Massive datasets often generated by complex simulation programs, required graphical visualization methods for best comprehension. Recent advances in multiscale visualization have been allowing the rendering to be done far more quickly and in parallel, making these visualization tasks practical.

Histograms (estimate the probability distribution for a certain numerical attribute given a set of objects), scatter plots (provide information on the relation between two numerical attributes and a discrete one), three-dimensional maps, dendrograms (a correlation analysis between attributes or objects) help in visualization.

In addition, some DM packages include *genetic algorithms* (optimization techniques based on the concepts of genetic combination, mutation, and natural selection), *sequential patterns* discovery (group objects with the same succession of given attribute values over a time period), *time-series similarity* (detect similar time series over a period of time), *Bayesian belief networks* (graphical models that encode probabilistic relationships among variables of interest, systems able to learn causal relationships), *neurofuzzy* systems (fuzzy inference systems that incorporate the learning and generalization abilities of neural works).

Genetic Algorithms. Optimization techniques use processes such as genetic combination, mutation, and neural selection in a design based on the concepts of natural evolution. Although genetic algorithms can be classified generally as an emerging science, it has a wide variety of applications in real-life situations.

The three main areas are:

(i) Optimization: Genetic algorithms can be used to automatically determine the optimal values for the variables that optimize the function.
(ii) Prediction: Genetic algorithms have been used at metalevel operations that are used to help optimize other data mining algorithms. For example, in optimizing weights in a neural network.
(iii) Simulation: Sometimes a specific problem is not well defined in terms of what the result is or whether one solution is better than the other. In such cases simulation can be done through genetic algorithms. GAs are systems that can solve a complex problem by arriving at pretty good solutions without a complete understanding of the problem.

Trend analysis. Time series data are records accumulated over time. For example, a company's sales, a customer's credit card transactions, and stock prices are all time series data. Such data can be viewed as objects with an attribute time. The objects are the snapshots of entities with values that change over time.

Sequencing. This technique helps identify patterns over time, thus allowing different analysis methods. The input data is a set of sequence called *data sequence*. Each data sequence is a list of transactions where each transaction is a set of items. A sequential pattern consists of a list of items. The problem is to find all sequential patterns with a user-specified minimum support, where the support of a sequential pattern is the percentages of data sequence that contain patterns.

Even if we wish to consider data mining tools like toolboxes of multiple techniques able to perform a complete data analysis, the reality is not yet so, the market offering presently only partially equipped products.

DM techniques are different one from another in terms of problem representation, parameters to optimize, accuracy, complexity, run time, transparency, and interpretability. Making a compromise between accuracy and complexity (by means of pruning techniques), enhancing the comprehension of derived patterns, and fitting to avoid the over fitting (a problem that appears when the model to be extracted is too complex with respect to the information provided in the learning set) are common features for all techniques.

2.3.2 Operational Aspects

The success of mining some data is induced by a list of factors:

The right tools. A distinctive feature of data mining software is the quality of its algorithms, the effectiveness of the techniques, and sometimes their speed. In addition, the efficiency of using the hardware, the operating system, the database resource, and the parallel computing influence the process. Moreover, it turns out that the particular set of tools useful in a given applications are highly dependent on the practical problem. Thus at the prototyping step, it is useful to have available a broad enough set of techniques so as to identify interesting applications. However, in the final product used for actual field implementation it is often possible to use only a small subset of the latter tools. Customizing data mining techniques to the application domain and using methods that are reliable means to the proposed goal may enhance the process of extracting useful information.

The right data. The data to be mined should contain information worth mining: consistent, cleaned, and representative for the application. Of course, it is useless to apply data mining to an invalid database with high measurement or estimation data errors, or to try to precisely estimate numerical outputs, which present a priori high noise. A data mining tool ideally explains as much information as is stored in the data that is mined (a derived model is strongly dependent on the learning set used), and sometimes it is not what is in the data that matters for an application (wrong attributes, wrong selected sample).

An important part of data mining result errors are due to uncertainties in modeling and generation of objects in certain databases discordant with the real probabilities of phenomena appearances in the system. That is why the data mining errors often do not have a meaning by themselves; rather they just provide a practical means of comparing efficiencies of different criteria applied to the same database.

The right people. Regardless of what many producers of data mining tools claim, data mining is not (yet) an "automatic" operation with little or no human intervention. On the contrary, the human analyst plays an important role, mostly in the areas of data selection and data/knowledge interpretation. The data miner should have an understanding of the data under analysis and

the domain or industry to which it pertains. It is more important for the mining process to embrace the problems of the application meant to solve, than to incorporate the hottest technologies in the data mining software.

The right application. Almost always a problem well posed is already a partially solved problem. It is important to clearly define the goals and choose the appropriate objectives so as to yield a significant impact on the underlying decision making process.

The right questions. An important issue is how the data mine structures a data analysis problem so that the right question can be asked, knowing how easy and useless it is to give the right answer to the wrong question.

The right sense of uncertainty. Data miners are more interested in understanding than accuracy or predictability per se. Often even the best methods of search will leave the data miner with a range of uncertainties about the correct model or the correct prediction.

2.3.3 The Need and Opportunity for Data Mining

Many of the techniques used by today's data mining tools have been around for many years, having originated in the artificial intelligence research of the 1980s and early 1990s. Yet these tools are only now being applied in the artificial to large-scale database systems. The confluence of several key trends is responsible for this new usage.

Widespread Deployment of High-Volume Transactional Systems. Over the past 15 to 20 years, computers have been used to capture detailed transaction information in a variety of corporate enterprises. Retail sales, telecommunications, banking, and credit card operations are examples of transaction-intensive industries.

These transactional systems are designed to capture detailed information about every aspect of business. Only five years ago, database vendors were struggling to provide systems that could deliver several hundred transactions per minute.

Information as a Key Corporate Asset. The need for information has resulted in the proliferation of data warehouses that integrate information from multiple, disparate operational systems to support decision making. In addition, they often include data from external sources, such as customer demographics and household information.

Widespread Availability of Scalable Information Technology. Recently, there has been widespread adoption of scalable, open systems-based information technology. This includes database management systems, analytical tools, and, most recently, information exchange and publishing through Intranet services.

These factors put tremendous pressure on the information "value chain." On the source side, the amount of raw data stored in corporate data warehouses is growing rapidly. The "decision space" is too complex; there is too much data and complexity that might be relevant to a specific problem. And on the sink side, the knowledge required by decision makers to chart the course of a business places tremendous stress on traditional decision-support systems. Data mining promises to bridge the analytical gap by giving knowledge workers the tools to navigate this complex analytical space.

2.3.4 Data Mining Tools and Techniques

Data mining applications can be described in terms of three-level application architecture. These layers include applications, approaches, and algorithms and models. These three layers sit on top of the data repository. We discuss these three levels in the following sections the characteristics of the data repository are addressed in the implementation section that follows.

Applications

We can classify data mining applications into sets of problems that have similar characteristics across different application domains. The parameterization of the application is distinct from industry to industry and application to application. The same approaches and underlying models used to develop a fraud-detection capability for a bank can be used to develop medical insurance fraud detection applications. The difference is how the models are parameterized – for example, which of the domain specific attributes in the data repository are used in the analysis and how they are used.

Approaches

Each data mining application class is supported by a set of algorithmic approaches used to extract the relevant relationship in the data: association, sequence-based analysis, clustering, classification, and estimation. These approaches differ in the classes of problems they are able to solve.

Association. Association approaches address a class of problems typified by a market-basket analysis. Classic market-basket analysis treats the purchase of a number of items (for example, the contents of a shopping basket) as a single transaction. The goal is to find trends across large numbers of transactions that can be used to understand and exploit natural buying patterns. This information can be used to adjust inventories, modify floor or shelf layouts, or introduce targeted promotional activities to increase overall sales, or move specific products. While these approaches had their origins in the retail industry, they can be applied equally well to services that develop targeted marketing campaigns or determine common (or uncommon) practices. In the

financial sector, association approaches can be used to analyze customers' account portfolios and identify sets of financial services that people often purchase together. They may be used, for example, to create a service "bundle" as part of a promotional sales campaign.

Association approaches often express the resultant item affinities in terms of confidence-rated rules, such as, "80% of all transactions in which beer was purchased also included potato chips." Confidence thresholds can typically be set to eliminate all but the most common trends. The results of the association analysis (for example, the attributes involved in the rules themselves) may trigger additional analysis.

Sequence-based analysis. Traditional market-based analysis deals with a collection of items as part of a point-in-time transaction. A variant of this problem occurs when there is additional information to tie together a sequence of purchases (for example, an account number, a credit card, or a frequent buyer/flyer number) in a time series. In this situation, not only may the coexistence of items within a transaction be important, but also the order in which those items appear across ordered transactions and the amount of time between transactions.

Rules that capture these relationships can be used, for example, to identify a typical set of precursor purchases that might predict the subsequent purchase of a specific item. In health care, such methods can be used to identify both routine and exceptional courses of treatment, such as multiple procedures over time.

Estimation. A variation on the classification problem involves the generation of scores along various dimensions in the data. Rather than employing a binary classifier to determine whether a loan applicant is a good or bad risk, this approach generates a credit-worthiness "score" based on a prescored training set.

Other techniques. Additional approaches used in conjunction with these and other analytical techniques include case-based reasoning, fuzzy logic, genetic algorithms, and fractal-based transforms. Fractal-based transforms (relatively new as data analysis tools) are interesting in that they are also used as aggressive, lossless data compression algorithms. Hence, there is the possibility that pattern-matching approaches based on these techniques could exploit substantially reduced dataset sizes to increase performance. Each of these has its own strengths and weaknesses in terms of the problem characteristics best addressed, discrimination capabilities, performance, and training requirements. The algorithms are often tunable using a variety of parameters aimed at providing the right balance of fidelity and performance.

Algorithms and Models

The promise of data mining is attractive for executives and IT professionals looking to make sense out of large volumes of complex business data. The

promise that programs can analyze an entire data warehouse and identify the key relationships relevant to the business is being pushed as a panacea for all data analysis woes. Yet this image is far from reality.

Today's data mining tools have typically evolved out of the pattern recognition and artificial intelligence research efforts of both small and large software companies. These tools have a heavy algorithmic component and are often rather "bare" with respect to users for interfaces, execution control, and model parameterization. They typically ingest and generate Unix flat files (both control and data files) and are implemented using a single-threaded computational model.

This state of affairs presents challenges to users that can be summed up in a sort of "tools gap." The gap caused by a number of factors require significant pre- and postprocessing of data to get the most out of a data mining application. Preprocessing activities include the selection of appropriate data subsets for performance and consistency reasons, as well as complex data transformations to bridge the representation gap. Postprocessing often involves subselection of voluminous results and the application of visualization techniques to provide added understanding. These activities are critical to effectively address key implementation issues such as:

Susceptibility to "dirty" data: Data mining tools have no higher-level model of the data on which they operate. They have no application-oriented (semantic) structure and as such, they simply take everything that are given as factual and draw the resulting conclusions. Users must take the necessary precautions to ensure the data being fed to the discovery tools. However, if the company has a good data cleansing process that cleans up data going into a data warehouse, then data mining tools benefit from this cleansing effort.

Inability to "explain" results in human terms: Many of the tools employed in the data mining analysis use complex mathematical algorithms that are not easily mapped into human terms; for example, they do not always generate "if-then" rules that use the original data attributes by name, so the ability of these systems to "explain" their results is minimal. Even with approaches such as decision trees and rule induction that are capable of generating information about the underlying attributes, the volume and format of the information may be unusable without additional postprocessing and/or visualization.

The data representation gap: Most of the source data for today's data mining applications resides in large, parallel relational database systems. The information is typically somewhat normalized and the attributes being used in a data mining application may span multiple tables. The data mining engines typically operate over a set of attribute "vectors" presented through a Unix flat file. Conditioning code must be used to provide the denormalized representation the tools need. Large central fact tables in data warehouses designed using star schema often combine denormalized data into one flat table.

Many of the tools are constructed in terms of the types of data elements with which that can work. Users may have to categorize continuous

variables or remap categorical variables. Time-series information may need to be remapped as well. For example, we might need to derive counts of the number of times a particular criterion was met in a historical database.

Although the 2-GB file is becoming less important with the advance of the 64-bit operating systems, many Unix implementations still have 2-GB file limitations. For flat file-based data mining tools, this limits the size of the datasets they can analyze, making sampling a necessity.

Parallel relational database systems store data that is spread across many disks and accessed by many CPUs. Current database architectures are such that result sets generated by the database engine are eventually routed through a single query coordinator process. This can cause a significant bottleneck in using parallel database resources efficiently. Because data mining applications are typically single-threaded implementations operating off Unix flat files, the process requires potentially large result sets to be extracted from the database.

Even if we are able to extract large datasets, processing then can be computation intensive. Although most data mining tools are intended to operate against data coming from a parallel database system, most have not been parallelized themselves.

This performance issue is mitigated by "sampling" the input dataset, which pose issues of its own. Users must be careful to ensure that they capture a "representative" set of records, lest they bias the discovery algorithms. Because the algorithms themselves determine those attributes that are important in the pattern matching, this presents a chicken-and-egg scenario that may require an iterative solution.

For algorithms that require training sets (classification problems), the training sets must adequately cover the population at large. Again this may lead to iterative approaches, as users strive to find reasonably sized training sets that ensure adequate population coverage.

Present-day tools are algorithmically strong but require significant expertise to implement effectively. Nevertheless, these tools can produce results that are an invaluable addition to a business's corporate information assets. As these tools mature, advances in server side connectivity, the development of business-based models, and user interface improvements will bring data mining into the mainstream of decision support efforts.

2.3.5 Common Applications of Data Mining

Data mining approach has a major advantage from the point of view of its applicability: almost all the domains of human activity may benefit from it, both the ones where a lot of data is already available and the ones where the data has to be simulated in order to extract some more profitable knowledge concerning the field. We mention further some particular broad domains of interest in the present data mining applications.

Market basket analysis refers to the process of examining point-of-scale data to identify affinities between products and services purchased by a customer. Data mining must deal in these applications with large volumes of transactional and spread data and must be performed in a time interval that will allow an organization to respond to market opportunity before competition does. Data mining techniques like association rules and sequential patterns discovery are involved in the automatic identification of important buying patterns, types of consumers exhibiting such patterns, customer characteristics that may be correlated to the consumer's choices.

Customer segmentation is the process of analyzing data about customers or general consumers to identify characteristics and behaviors that can be exploited in the market place. Clustering, statistical analysis, deviation detection, and modeling are implicated in reducing the customer attrition phenomenon, i.e., the loss of customers (searching for customers who exhibit characteristics typical of someone who is likely to leave for a competing company), or in target marketing (attracting other customers, identification of the risk associated with insurance, etc.).

Fraud detection. Data mining applications have demonstrated their benefits in the areas where many actions (transactions) are undertaken, making the respective system vulnerable to fraud: credit card services, telecommunications, computer systems, etc.

Detection of patterns in text, image, on the World Wide Web are broadly extensive areas of DM applications due to the impressive of information available: finding association amongst the keywords labeling items in a collection of textual documents, recognizing actions in video image sequences, helping users locate desired information in the Web, and so on

Medical diagnosis through means of data mining are intended to be helpful tools that can improve the physicians' performance and make the diagnosis process more objective and more reliable. From the descriptions of the patients treated in the past for which the final diagnosis were verified, diagnosis rules may be automatically derived by means of clustering machine learning, association rules, although the technology is not widely accepted in medical practice, encountering a resistance of the physicians to new diagnostic technology.

2.3.6 What about Data Mining in Power Systems?

Why would data mining tools be useful in the power system field? Like many other application areas, the power system field is presently facing an explosive growth of data. In power systems, irrespective of the particular application, there are three main sources of data: (i) field data, collected by various devices distributed throughout the system, such as digital records; (ii) centralized data archives, such as those maintained by control center SCADA systems, (iii) data from simulations, carried out in planning or operation environments.

In a power system there are a few DM-related aspects: large-scale character of power systems (thousands of state variables), temporal (from milliseconds to minutes, hours, weeks, years), and statistical nature of data, existence of a discrete (e.g., events such as topology changes or protection arming) and continuous (analog state variables) information mixture, necessity of communication with experts through means of visualization, online operation time restrictions for fact decision making existence of uncertainty (noise, outliers, missing information).

Engineers trying to solve power system-related problems should look at the whole toolbox of data mining methods and not hesitate to combine different techniques to yield a full, practical solution. Data selection step may be performed with a decision tree, a clustering approach, or a correlation analysis, and later on, the result may serve as input for other supervised techniques, possibly with the problem decomposed into simpler subproblems.

There are three dimensions along which data mining may complement classical system theory-oriented methods for power systems.

Computational efficiency. By using synthetic information extracted by DM, instead of numerical methods, much higher speed may be reached for real-time decision making. Further, in terms of data requirements, DM may require only significant and/or available input parameters database, instead of a full description of the system model.

Anticipative physical insight. The present-day practice generally handles new problems, though some undesirable consequences have already been observed on the system. Carrying out DM studies will allow the engineers to have a more anticipative view on potential problems.

Management of uncertainties. The behavior of a power system will always have some unexpected experiences (e.g., a relay that misoperated, an operator that did not behave as expected, a set point that was different from prescriptions, a load that was modeled inappropriately). DM copes with this problem by making use of more simulations carried out by relaxing assumption on the dynamic models used.

2.4 Data Warehousing and OLAP

2.4.1 Data Warehousing for Actuaries

Competition in the insurance industry has never been stronger. The emergence of bancassurance, the recent frenzy of insurance company demutualizations and mergers, and the potential of the Internet all add to this. To stay ahead in this market, companies must understand their business, in particular: Knowing who their customers are, understanding the profile of their business, identifying the most profitable customers and products, and understanding what makes producers successful. Insurance business is complex, characterized

by large volumes of data, dating back many years, with complex data relationships. This data is often fragmented throughout the organization, stored on multiple systems, with the resulting information retained in small analytical niches. To address these data problems many companies have turned to data warehouses. Promising to be a single version of reality, a data warehouse has the obvious appeal of collecting all the data and storing it in a single place – somewhere it can be queried and manipulated, without affecting the day-to-day running of the business. But it is more than that, the data is structured in a way to optimize analytical applications, definitions are added to data elements to ensure a common understanding, and front-end tools have been developed that can efficiently drill into, manipulate, and distribute results. Data warehouses are now central to the systems strategy for companies in many industries. However, the life insurance industry has been slow to realize the full potential of this technology. The reasons for this are numerous, but at the forefront is the complexity of the business and the data relationships. Some companies have explored marketing data marts and reporting marts, but more complex analysis such as earnings by source and experience analysis is still rare.

2.4.2 Data Warehouse Components

The data warehouse will typically consist of several components. Data originates from various source administration systems, extracted as a series of snapshots at regular time intervals; typically monthly or quarterly for financial analysis, or more frequently for marketing analysis.

The ETL (Extraction, Transformation, and Loading) tool cleans and transforms the data, so that it meets the requirements of the warehouse; information about this process is stored as metadata. This metadata, together with business metadata (business rules and definitions), is available to end users of the system.

Data stored on the source systems is unlikely to be in the format required. Transformation can be as simple as converting "0" to "male" and "1" "female," or very complex involving multiple data fields to be evaluated simultaneously and logic applied to determine the desired value. This stage is critical as it provides users with confidence in the data they are using. It also provides system management benefits such as:

- An audit trail and documentation of the extract routine
- A formal loading process so that reloads are possible
- Better automation than ad hoc loading of data.

Data may also be fed to and from third-party systems, such as valuation systems, to make use of their specialist functions. OLAP or data cubes provide a simple and effective way to view data.

Data warehousing for actuaries cube can be thought of as the dimensions of the data – how the data is stored and viewed, such as by product line,

time period, currency. The content of the cube is the item being measured, such as premium, reserve, or sum assured. It is easy to visualize how data is accessed, by identifying the content at the intersection of the selected dimensions. But here the analogy breaks down; data views are not limited to only three dimensions as physical cubes are. In addition, intelligence can be built in the dimensions to help the user by identifying natural drill down paths, such as product lines and time periods. Data is usually presented to users in the form of pivot tables, and may also consist of preformatted reports, Web portal, or graphical interface. Data warehousing tools bring a number of natural strengths to the sort of analysis that is required for insurance. Consistency is clearly enhanced with the introduction of a data warehouse. If the administration data is on multiple systems it is an opportunity to bring it together on one system, with consistent definitions for status, gender, and other demographic information that would be common across systems, but not necessarily stored consistently. Also, having assets and liabilities together on a single system allows consistent and controlled analysis of both sides of the balance sheet. Multiple hierarchies may be used to meet the requirements of various reports and reporting bodies. These will be consistent at the lowest level of detail; reconciliation is then simple with the help of metadata and drill down. Similarly demographic groupings can be predefined. With the ability to drill down into the demographic groupings reports should be clearer and more consistent.

2.4.3 Management Information

Most insurance companies have multiple sources of management information. The data warehouse provides an opportunity to combine these and at the same time ensure consistency between subject areas. These management information reports may include:

- Business profiles such as new business and in-force policy reports.
- Movement analysis.
- Experience studies.
- Financial reporting.

Even though financial reporting is basically the same every year or every quarter, it is often at a time when long hours are worked with last-minute surprises in the results. A data warehouse can strengthen any process and remove redundant processes, such as multiple data extracts. Ultimately, financial reporting consists of a few results presented in different ways and tested for trends and reasonableness. With a data warehouse these can prepared in a seamless fashion and can eliminate a lot of time in reconciliation between the various reports that have essentially the same numbers aggregated in different ways. This also reduces the chance of errors occurring in transferring data and helps simplify audit and process documentation. Previously, actuaries would have spent a significant amount of their time collecting and collating data.

With the data warehouse, actuaries can better spend their time understanding the impact of the results.

2.4.4 Profit Analysis

Profit analysis is more than the ability to present details of the published statement. Profit analysis should highlight the elements that make up the profit. Profit should be understood in terms of the source factors that contribute to it; the lapse contribution, for example, is not just the number of lapses for in the year, but also the impact on the reserves, mortality, and expenses.

The actuarial methods used for analyzing profit, when deployed in a data warehouse, allow profit to be viewed for any policy attribute such as product or product line, geographical sales area, premium range, etc. by simply limiting on the appropriate dimensions. Furthermore, with this implementation approach it is possible to drill in and constrain on any of the product attributes to give rich and informative reports on the business. This flexibility of presentation and detail can be extremely useful when communicating a difficult subject.

2.4.5 Asset Liability Management

Regular ALM reporting can be made quickly available with the use of the data warehouse. Just as with profit analysis the additional flexibility provided by the use of dimensional drill-down can really help communication. Insurance companies tend to have a significant amount of unstructured duplication of effort and data. By using a data warehouse and structuring data in way that suits analysis, organizations will quickly discover that reports and analysis become interrelated. Time and money are saved, as additional reconciliation is not required. More importantly, from marketing analysis, such as customer behavior to financial analysis, the depth of understanding is increased with the ability to quickly manipulate and share information.

Scorecards. Balanced scorecards and financial scorecards are becoming more common in organizations. A scorecard is basically a set of performance dials for measuring strategic health and performance. This approach can be used to translate vision, strategy, and tactics into concrete financial performance targets and measures. And in turn, for using these targets and measures for communication of what is important to the business.

Data Mining. Data mining in insurance companies has been used primarily for marketing purposes and also to assist with underwriting in avoiding antiselection. But for many insurance companies data mining is not practical; the quality of the data on the source system is likely to swamp any useful discovery. However, the data warehouse provides a natural platform on which to base any data mining activities. Any patterns that are discovered by the data

mining tool are likely to require some justification. Often these are counterintuitive and time will have to be spent examining this before management will be persuaded to take action. However it could be argued that this is precisely the purpose of data mining.

2.5 Data Mining and OLAP

In the section that follows we give perspective of data mining such as: Why do we need data mining? How are other DM techniques are used? As an example of using data mining and OLAP, we have introduced in this section the definition of OLAP, which is short for Online Analytical Processing, a category of software tools that provide analysis of data stored in a database. OLAP tools enable users to analyze different dimensions of multidimensional data. For example, OLAP provides time series and trend analysis views. OLAP covers, including multiple hierarchies, rules analytical operations and the difference between OLAP and data mining.

2.5.1 Research

1. History Perspective – The Relational Model

Transaction processing systems. This is where clients store messages in a DBMS storing the data by using a number of files such as sequential files, indexed sequential access methods, variable sequential access methods, hierarchical databases and network databases. OLTP are good for putting data into databases.

The relational databases management system stores all sorts of business data by using the sequential files or indexed sequential access method or variable sequential access method, or hierarchical databases and network databases. Also it increases the electronic devices that enter all the data into database storage, such as a barcode scanner, EPOS, etc. It has become a useful accessory, which is cheaper and more powerful.

2. DBMS Problems

DBMS has given the access to store the data but there is no analysis to analyze the data. Because there is so much data, it would take a long time to find a specific data we require. With analysis there are all sorts of data that could be used such as decision support. Without the analysis, a new user would know nothing about the commercial business application and paralleled principles.

This is where data mining has become increasingly important to the organization, giving them more of an advantage. Data mining is the use of software that provides tools and applications (data analysis, decision support, and automation etc.) to gather reports and analyze the information, such as

- Survey results analysis
- Inappropriate practices
- Fraud detection
- Manufacturing process analysis
- Risk analysis and management
- Market and sales analysis
- Scientific data analysis
- Text document analysis

It is also used to collect all the necessary information and store it in electronic format for the future. Data mining involves capturing, storing, gathering, and analyzing the company information. There is a risk of losing such amounts of important data; hence the company will have to consider being careful about giving access or changing the system.

3. Data Mining

The definition of data mining is:

> "*Data mining is the search for relationships and global patterns that exist in large database but are 'hidden' among the vast amount of data, such as a relationship between patient data and their medical diagnosis. These relationships represent valuable knowledge about the database and the objects in the database and, if the database is a faithful mirror, of the real world registered by the database*".
>
> Marcel Holeshemier & Arno Siebes (1994)

Who needs Data Mining?

Every business needs data mining because companies are looking for new ways to access and to allow end users access to the data they need for making decisions, serving customers, and gaining the competitive edge.

We need data mining to collect data and allow us to analyze it. If it is not analyzed at that time the collected data could be important in the near future. As a database grows, the ability to support decisions by processing using traditional query languages is not feasible.

Data Mining Process

A data mining project consists of a life cycle that has six phases. It is not strict that these phases should be in order. It can either move backward or forward in between the phases, as they are required, in each phase it becomes the outcome of the task and is performed in the next phase. This indicates the most important and frequent dependencies between phases. The data mining process continues after a solution has been deployed. This has become new

and often focuses more on business questions that subsequent data mining process will benefit from the experience of previous ones.

The steps below help to explain each stage of the data mining process:

- Business understanding

This focuses on understanding the project objectives and requirements from a business perspective and finds a data mining problem of definition. Then it is so designed to achieve the objectives.

- Data understanding

This subject deals with understanding the data before previous data is collected and proceeds with activities; so it will be familiar with the data to identify data quality problems, discover the data, and then start proceeding.

- Data preparation

The data preparation phase covers all activities to find the data that will be fed into modeling tools, from the initial raw data. In this phase the tasks are likely to be performed multiple times, and it is not required to do anything in sequence order. Tasks include table and record, while attribute selection includes transformation and cleaning of data for modeling tools.

- Modeling

This phase has various modeling techniques that are selected and applied, so their parameters are adjusted to optimal values. Sometimes, there are several techniques for the same data mining problem type. There are some techniques that have specific requirements in the form of data. So they step back to the data preparation phase, which is often needed.

- Evaluation

By this stage we would have built a model(s) for the project so that it appears to have high quality, hence we get data from analysis perspective. Before going to further phases of the model, it is important to thoroughly evaluate the model and review the steps to the end to construct the model – this can help in properly achieving the business objectives. This is a key objective to determine if there are some important business issues that need to be considered. At the end of this phase, this decision will find the use of the data mining results.

- Deployment

Creating the model is not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the consumer can use it. On the requirements, this phase can be as simple as generating a report or as complex

as implementing a repeatable data mining process. In many cases, it will be the customer, and not the data analyst, who will carry out the deployment steps. Therefore if the analyst does not carry out this deployment, it is important to make an effort for the customers to understand the action that is to be carried out in order to actually make use of the created models.

Main points of Data mining for the organization

- They will have large volumes of data
- Employees need to understand data
- There is a need to exist in a competitive market
- Data mining components (discovery and exploitation)
- Data mining discovers relationships between data

Data Mining Components (Discovery and Exploitation): Data mining makes discovery of facts easier as they are represented as information data. Then follows the exploitation or use of those facts for problem solving.

Data Exploration: This involves preparing the data to get a better pattern discovery and also to validate the result of the data preparation. But sometimes this involves examining the statistics (minimum, maximum, average, etc.) and the frequency distribution of individual data fields. Finally this involves the field versus field graphs to understand the dependency between fields.

Pattern Discovery: This is the stage of applying the pattern discovery algorithm to generate patterns. This allows business users to interact with and to impart their business knowledge to the discovery process. This stage also involves analyzing for understanding discovered patterns to predict the propensity of the business event, and for verification against an independent dataset.

Data Mining discovers relationships between data: All new data can be added easily to existing data so that if there is a relationship between any of them a solution may be worked out. For example if a woman asks her husband to buy apples in the supermarket, it is likely that he will want to buy soft drink for himself at the same time. So the management may suggest that the soft drink section be placed quite near to the baby goods shelves to encourage sales. This plan of action is used to promote most products in large shops.

Making choices – A decision has to be made as to which is a suitable choice from various alternatives.
Making predictions – A prediction is a choice about what may happen in the near future.
Improving processes – To improve the system, the current system must be updated and improved.
Mining For Golden Nuggets of Data.

Issues in Data mining

There are some problems. Data mining systems rely on databases to supply the raw data for input. This can cause problems in that the database tends to be dynamic, incomplete, noisy, and large. Then there are other problems that have arisen as a report of the adequacy and relevance of the information stored.

Limited Information

Sometimes the database is designed differently for the data mining; therefore the problem has to be solved before access to the information can be given. Undecided data causes problems; this can be because some attributes have useful knowledge about the application. This is not present in the data that may be impossible to discover the knowledge about a given domain. Databases are usually muddled up by errors, so it cannot be the correct way the data is contained from the originally inputted data.

Missing data can be treated by discovery systems in a number of ways such as:

- Simply disregarding missing values
- Omitting the corresponding records
- Inferring missing values from known values
- Treating missing data as a special value to be included additionally in the attribute domain.
- Averaging over the missing values using basic techniques.

Uncertainty refers to the strictness of the error and the noise in the data. Databases tend to be large and dynamic in that their contents change as information is added, modified, or removed.

Data mining Applications

Data mining in Government

- Develop advanced pattern recognition
- Identify problem in technologies
- Improve the project of their techniques

Data mining in Energy

- Identify loyal customer
- Fraud detection in energy industry
- Predict industry specific

4. OLAP

There are two technologies that are related to data mining: data warehouse and OLAP.

Data Warehouses

Here the data is collected and stored in the data warehouse. Data warehouse is the relational database management system, but it is specifically designed for a transaction processing system. These warehouses contain millions of pieces of information about customer's needs and distribution decisions. Data warehouse uses the data to analyze business needs and to make the decisions. They make amounts of data that span over many years. A data warehouse is not a transactional database.

Transactional databases use data to operate business needs. For instance, if an organization wants to look up data about 60–90 days old, they can. They are built for speed and efficiency to keep the company's day-to-day operations moving fast, unlike data warehouses, which runs very slowly.

This is where OLAP (On Line Analytical Processing) introduces the tools to analyze data warehouse information. We need data mining to collect data and to analyze it, or if not analyzed, collected. This could be important in the near future. As databases grow, the ability to support decision by processing traditional query languages is no longer feasible.

What is OLAP?

OLAP provides the users with multidimensional database to generate on-line description, or it compares the "views" of data and other analytic queries. OLAP gives the answers to multidimensional business questions quickly and easily. OLAP technology provides facts and efficient access to summarized data. Also OLAP is able to give control over global views of the business.

OLAP technology can be applied to:

- Sales and marketing analysis
- Financial reporting
- Quality tracking
- Profitability analysis
- Manpower and pricing applications
- Our unique data discovery needs

The users have found easy-to-understand graphical information that can be presented simply and acted upon quickly. Like for instance, when the clients have problems with their data mining process, they need to respond quickly for their requirement.

Relational OLAP approach builds such a system. Relational databases are used to build and query these systems. To analyze the query using in SQL

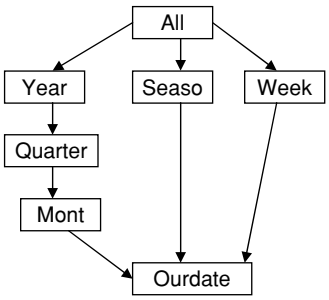

Fig. 2.4. Multiple dimensional hierarchies

might not be efficient to execute; therefore another away of using multidimensional database techniques has been applied to decision support applications. Data is stored in multidimensional structure, which is a more easy way of expressing the enterprise data and is more suited for analysis.

Multiple hierarchies

A dimension can consist of multiple hierarchies as shown in Fig. 2.4. This enables a consistent representation of the information in, e.g., data warehouse, yet provides flexibility to see the information using different perspectives.

There are three reasons why it is beneficial to create multiple hierarchies.

- Sharing a single key in the fact table. It can create a single key to represent both hierarchies in the fact table, thus reducing the fact table size and complexity.
- Usability. Multiple hierarchies in a single dimension provide better usability for the end user. If the front-end application supports multiple hierarchies logically grouped together.
- Sharing of aggregates. Suppose we have three hierarchies. One hierarchy is year, month, day, date; the second is season and date; and third is week and date. The three hierarchies share the aggregates that are built on the year level, thereby reducing the processing time necessary to create the same aggregate second and third times.

OLAP Rules

Originally, 12 OLAP rules were defined to characterize this technology. OLAP products tend to conform to these. The 12 rules are as follows:

- Multidimensional model
- Transparency of the server
- Accessibility
- Stable access performance

- Client server architecture
- Generic dimensionality
- Management of data
- Multiuser
- Operation on dimension
- Intuitive manipulation of data
- Flexible posting and editing
- Multiple dimensions and levels

Today these are the 12 main rules that have been expanded to 18 major rules and up to 300 in total.

Example of BI (Business Intelligence) Technology

By using ad hoc query application the user can have access to information on demand; this means that what they ask for is what they will get. For example a user creates and executes an ad hoc query by asking the question, "How much revenue was generated by each customer during this year?" The results from the query would ask the customer name and give the revenue for the year selected.

Then the customer asks a series of questions, "How much revenue was generated this year?" "Which customer generated the most revenue for the company?" "And which customer generated the least revenue for the company?"

BI technology could not identify the unusual patterns or reveal unusual relationships. This is the next stage where the OLAP comes in. OLAP analyses the summary and details the information. They could answer all the questions above. The user creates and performs an OLAP analysis that provides answers to the questions "What was the revenue for each quarter of this year by geographic region and customer?" The result from this analysis would contain geographic region, customer name, revenue, and quarter selected.

But the user needs to understand how to navigate the data, which must direct these processes. OLAP can only highlight the patterns within the data that was requested. So the BI technology will not identify unusual or reveal hidden relationships. The next stage is where the data mining comes in.

In BI technology data mining can extract the hidden and useful information from the data. Data mining discovers the hidden trends and patterns in large amounts of data. There are various techniques that can be deployed, each serving a specific purpose and varying amounts of user involvement.

2.5.2 Data Mining

Neural networks –detect potential fraudulent credit card transactions.
Induction – understands the relationship that exists, e.g., when people buy nappies, they also buy beer 50% of the time.

Statistics – requires highly skilled mathematicians to build and interpret the results.

Visualization – displays the data in a graphical map for the user to identify trends, patterns, and relationships. Images produced provide another perspective of data relationship. Visualization is often incorporated in data mining application. Ad hoc query applications stated the values that exist within a database, while OLAP provides users with understanding and gives more detailed information.

Data mining digs deeper and provides users with knowledge through the discovery of hidden trends and relationships. By using ad hoc query, OLAP application and data mining becomes powerful, it provides users with knowledge about the data that is analyzed and enables action on the knowledge that has been found.

Analytical Operations

The end users must get the data for analytical purposes by applying analytical operations such as ratios, cumulative totals, trends, and allocations across dimensions and hierarchical levels. OLAP functionally is described by dynamic multidimensional analysis of:

Consolidated data supporting the end user analytical navigational activities including

- Calculations and modeling applied across dimensions, through hierarchies, and across members
- Trend analysis over sequential time periods
- Slicing subsets for on-screen viewing
- Drill down to the deeper levels of consolidation
- Each through to underlying detail data
- Rotation to new dimensional comparisons in the viewing area

OLAP and Data Mining

OLAP is user driven, the analyst generates a hypothesis and uses OLAP to verify, e.g., "people with high debt are bad credit risks."

Data mining tool generates a hypothesis – Tool performs exploration, e.g., find risk factors for granting credit.

It also discovers new patterns that analysts did not think of, e.g., debt-to-income ratio. OLAP and DM need each other; we have mentioned this in the example about (BI Technology). OLAP and data mining should not remain separate components of decision support; therefore they should be fully merged and they really need each other. When dealing with OLAP and data mining it shows that decision support applications must consider data mining within multiple dimensions and that OLAP system needs to focus on discovery as much as on access. OLAP and data mining must work together within the framework of any theory to avoid getting incorrect results.

OLAP and OLTP

A major issue in information processing is how to process increasingly larger databases, containing increasingly complex data, without sacrificing response time. The client/server architecture gives organizations the opportunity to deploy specialized servers, which are optimized for handling specific data management problems. Until recently, organizations have been trying to target relational database management systems (RDBMSs) for the complete spectrum of database applications. It is however apparent that there are major categories of database applications that are not suitably serviced by relational database systems.

OLAP Example

An example of OLAP database may comprise sales data that has been aggregated by region, product type, and sales channel. A typical OLAP query might access a multigigabyte/multiyear sales database in order to find all product sales in each region for each product type. After reviewing the results, an analyst might further refine the query to find sales volume for each sales channel within region/product classifications. As a last step the analyst might want to perform year-to-year or quarter-to-quarter comparisons for each sales channel. This whole process must be carried out on-line with rapid response time so that the analysis process is undisturbed. OLAP queries can be characterized as on-line transactions, which:

- Access very large amounts of data, e.g., several years of sales data.
- Analyze the relationships between many types of business elements, e.g., sales, products, regions, channels
- Involve aggregated data, e.g., sales volumes, budgeted dollars, and dollars spent.
- Compare aggregated data over hierarchical time periods, e.g., monthly, quarterly, or yearly
- Present data in different perspectives, e.g., sales by region vs. sales by channels by product within each region.
- Involve complex calculations between data elements, e.g., expected profit as calculated as a function of sales revenue for each type of sales channel in a particular region.
- Are able to respond quickly to user requests so that users can pursue an analytical thought process without being stymied by the system.

Comparison of OLAP and OLTP

OLAP applications are quite different from on-line transaction processing (OLTP) applications, which consist of a large number of relatively simple transactions. The transactions usually retrieve and update a small number of records that are contained in several distinct tables. The relationships between the tables are generally simple.

A typical customer order entry OLTP transaction might retrieve all of the data relating to a specific customer and then insert a new order for the customer. Information is selected from the customer, customer order, and detail line tables. Each row in each table contains a customer identification number, which is used to relate the rows from the different tables. The relationships between the records are simple and only a few records are actually retrieved or updated by a single transaction.

The difference between OLAP and OLTP has been summarized as while OLTP servers handle mission-critical production data accessed through simple queries, OLAP servers handle management-critical data accessed through an iterative analytical investigation. Both OLAP and OLTP have specialized requirements and therefore require special optimized servers for the two types of processing.

OLAP database servers use multidimensional structures to store data and relationships between data. Multidimensional structure contains aggregated data relating elements along each of the dimensions. For example, a single cell may contain the total sales for a given product in a region for a specific sales channel in a single month. Multidimensional databases are a compact and easy-to-understand vehicle for visualizing and manipulating data elements that have many inter-relationships.

OLAP database servers support common analytical operations, including consolidation, drill-down, and slicing and dicing.

- Consolidation - involves the aggregation of data such as simple roll-ups or complex expressions involving inter-related data. For example, sales offices can be rolled up to districts and districts rolled up to regions.
- Drill-down - OLAP data servers can also go in the reverse direction and automatically display detail data, which comprises consolidated data. This is called drill-downs. Consolidation and drill-down are an inherent property of OLAP servers.
- Slicing and Dicing - Slicing and dicing refer to the ability to look at the database from different viewpoints. One slice of the sales database might show all sales of product type within regions. Another slice might show all sales-by-sales channel within each product type. Slicing and dicing is often performed along a time axis in order to analyze trends and find patterns.

OLAP servers have the means for storing multidimensional data in a compressed form. This is accomplished by dynamically selecting physical storage arrangements and compression techniques that maximize space utilization. Dense data (i.e., data exists for a high percentage of dimension cells) is stored separately from sparse data (i.e., a significant percentage of cells are empty). For example, a given sales channel may only sell a few products, so the cells that relate sales channels to products will be mostly empty and therefore sparse. By optimizing space utilization, OLAP servers can minimize physical storage requirements, thus making it possible to analyze exceptionally large amounts of data. It is also possible to load more data into computer

memory, which helps to significantly improve performance by minimizing physical disk I/O.

In summary, OLAP servers logically organize data in multiple dimensions, which allows users to quickly and easily analyze complex data relationships. The database itself is physically organized in such a way that related data can be rapidly retrieved across multiple dimensions. OLAP servers are very efficient when storing and processing multidimensional data. RDBMSs have been developed and optimized to handle OLTP applications. Relational database designs concentrate on reliability and transaction processing speed, instead of decision support need. The different types of server can therefore benefit a broad range of data management applications.

2.6 Summary

Data mining is beginning to contribute research advances of its own, by providing scalable extension and advances to work in associations, ensemble learning, graphical models, techniques for on-line discovery, and algorithms for the exploration of massive and distributed datasets.

Advances in data mining require supporting a) single investigators working in data mining and the underlying research domains supporting data mining; b) interdisciplinary and multidisciplinary research groups working on important basic and applied data mining problems; and c) the appropriate test beds for mining large, massive, and distributed datasets. Appropriate privacy and security models for data mining must be developed and implemented.

There are three dimensions along which data mining may complement classical system theory-oriented methods for power systems. They are computational efficiency, anticipative physical insight, and management of uncertainties that are discussed earlier in this section.

Overall in this section more information is added to data mining and OLAP. There are many differences between data mining and OLAP and they have been used for the organization business (e.g., BT technology). Data mining and OLAP should work together; therefore the data has given faster information by using their techniques tools. Data mining has a lot of potential; it is increasing every year and is becoming very useful for organizations. It is fast and efficient along with OLAP. It has diversity in the field of application, which becomes more efficient in the database system. Now the estimated market for the data mining in the United States in $500 million.

2.7 Review Questions

1. What are the recent research achievements in data mining?
2. State some of data mining application areas.
3. Write about the various trends in the current scenario that affect data mining.

4. Give some of the research challenges in data mining area.
5. Define the KDD process.
6. Explain in detail about the process involved in data mining.
7. What are the major essential elements in data mining?
8. Who is a data miner and write on the approach of data miners in mining.
9. What factors influence the success of data mining?
10. Explain data mining in power systems with suitable example.
11. What are the components of data warehouse and data mining?
12. Define OLAP.
13. Give details on multidimensional hierarchies used in mining.
14. State the OLAP rules.
15. Explain about business information technology using OLAP.
16. Write short notes on OLAP and OLTP.
17. Compare and contrast OLAP and OLTP.