A

Seminar report

On

Data Mining

Submitted in partial fulfillment of the requirement for the award of degree
Of Computer Science

SUBMITTED TO:

www.studymafia.org

SUBMITTED BY:

www.studymafia.org

Preface

I have made this report file on the topic **Data Mining**, I have tried my best to elucidate all the relevant detail to the topic to be included in the report. While in the beginning I have tried to give a general view about this topic.

My efforts and wholehearted co-corporation of each and everyone has ended on a successful note. I express my sincere gratitude to ……..who assisting me throughout the prepration of this topic. I thank him for providing me the reinforcement, confidence and most importantly the track for the topic whenever I needed it.

Acknowledgement

I would like to thank respected Mr…….. and Mr. ……..for giving me such a wonderful opportunity to expand my knowledge for my own branch and giving me guidelines to present a seminar report. It helped me a lot to realize of what we study for.

Secondly, I would like to thank my parents who patiently helped me as i went through my work and helped to modify and eliminate some of the irrelevant or un-necessary stuffs.

Thirdly, I would like to thank my friends who helped me to make my work more organized and well-stacked till the end.

Next, I would thank Microsoft for developing such a wonderful tool like MS Word. It helped my work a lot to remain error-free.

Last but clearly not the least, I would thank The Almighty for giving me strength to complete my report on time.

Content

- Introduction
- Scope of Data Mining
- Data Mining Process
- Data mining Architecture
- Data mining Techniques
- Application of Data Mining
- Advantages Data Mining
- Disadvantages Data Mining
- Conclusion
- Reference

Introduction

Data mining, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. When implemented on high performance client/server or parallel processing computers, data mining tools can analyze massive databases to deliver answers to questions such as, "Which clients are most likely to respond to my next promotional mailing, and why?"

This white paper provides an introduction to the basic technologies of data mining. Examples of profitable applications illustrate its relevance to today's business environment as well as a basic description of how data warehouse architectures can evolve to deliver the value of data mining to end users.

The Scope of Data Mining

Data mining derives its name from the similarities between searching for valuable business information in a large database — for example, finding linked products in gigabytes of store scanner data — and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

- **Automated prediction of trends and behaviors**. Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data — quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

- **Automated discovery of previously unknown patterns**. Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

Data mining techniques can yield the benefits of automation on existing software and hardware platforms, and can be implemented on new systems as existing platforms are upgraded and new products developed. When data mining tools are implemented on high performance parallel processing systems, they can analyze massive databases in minutes. Faster processing means that users can automatically experiment with more models to understand complex data. High speed makes it practical for users to analyze huge quantities of data. Larger databases, in turn, yield improved predictions.

Databases can be larger in both depth and breadth:

- **More columns**. Analysts must often limit the number of variables they examine when doing hands-on analysis due to time constraints. Yet variables that are discarded because they seem unimportant may carry information about unknown patterns. High performance data mining allows users to explore the full depth of a database, without preselecting a subset of variables.

- **More rows**. Larger samples yield lower estimation errors and variance, and allow users to make inferences about small but important segments of a population.

A recent Gartner Group Advanced Technology Research Note listed data mining and artificial intelligence at the top of the five key technology areas that "will clearly have a major impact

across a wide range of industries within the next 3 to 5 years."2 Gartner also listed parallel architectures and data mining as two of the top 10 new technologies in which companies will invest during the next 5 years. According to a recent Gartner HPC Research Note, "With the rapid advance in data capture, transmission and storage, large-systems users will increasingly need to implement new and innovative ways to mine the after-market value of their vast stores of detail data, employing MPP [massively parallel processing] systems to create new sources of business advantage (0.9 probability)."3

The most commonly used techniques in data mining are:

- **Artificial neural networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

- **Decision trees**: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) .

- **Genetic algorithms**: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.

- **Nearest neighbor method**: A technique that 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 ³ 1). Sometimes called the k-nearest neighbor technique.

- **Rule induction**: The extraction of useful if-then rules from data based on statistical significance.

Many of these technologies have been in use for more than a decade in specialized analysis tools that work with relatively small volumes of data. These capabilities are now evolving to integrate directly with industry-standard data warehouse and OLAP platforms. The appendix to this white paper provides a glossary of data mining terms.

Data Mining Processes

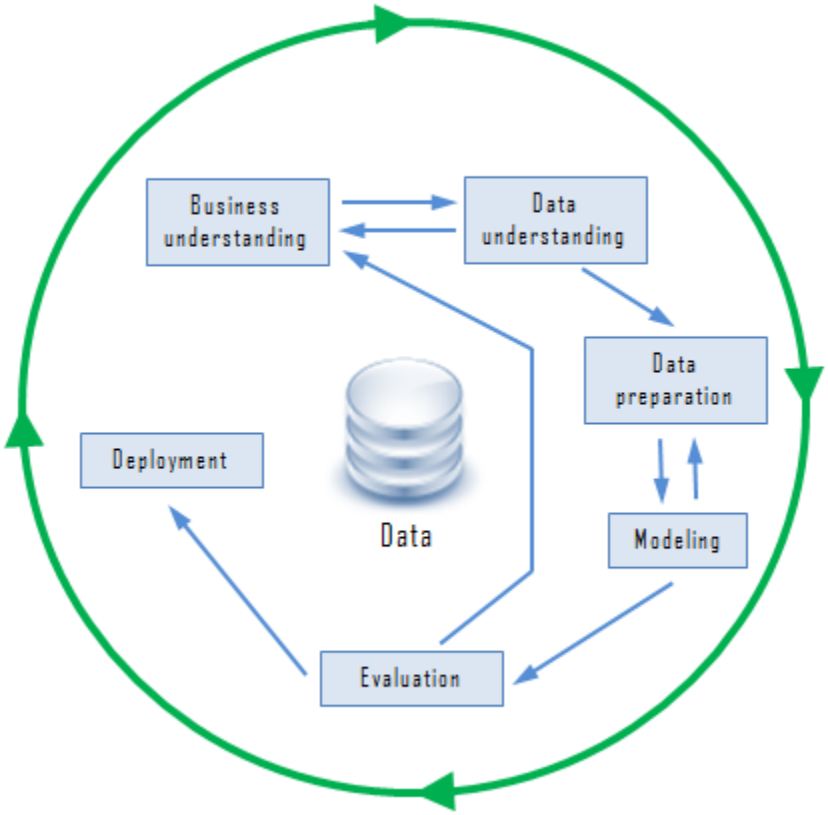
Data mining is a promising and relatively new technology. Data mining is defined as a process of discovering hidden valuable knowledge by analyzing large amounts of data, which is stored in databases or data warehouse, using various data mining techniques such as machine learning, artificial intelligence(AI) and statistical.

Many organizations in various industries are taking advantages of data mining including manufacturing, marketing, chemical, aerospace… etc, to increase their business efficiency. Therefore the needs for a standard data mining process increased dramatically. A data mining process must be reliable and it must be repeatable by business people with little or no knowledge of data mining background. As the result, in 1990 a cross-industry standard process for data mining (CRISP-DM) first published after going through a lot of workshops, and contributions from over 300 organizations.

Let's examine the cross-industry standard process for data mining in greater detail.

The Cross-Industry Standard Process for Data Mining (CRISP-DM)

Cross-Industry Standard Process for Data Mining (CRISP-DM) consists of six phases intended as a cyclical process as the following figure:



Cross-Industry Standard Process for Data Mining (CRISP-DM)

Business understanding

In the business understanding phase:

- First, it is required to understand business objectives clearly and find out what are the business's needs.
- Next, we have to assess the current situation by finding about the resources, assumptions, constraints and other important factors which should be considered.
- Then, from the business objectives and current situations, we need to create data mining goals to achieve the business objectives within the current situation.
- Finally, a good data mining plan has to be established to achieve both business and data mining goals. The plan should be as detailed as possible.

Data understanding

- First, the data understanding phase starts with initial data collection, which we collect from available data sources, to help us get familiar with the data. Some important activities must be performed including data load and data integration in order to make the data collection successfully.
- Next, the "gross" or "surface" properties of acquired data needs to be examined carefully and reported.
- Then, the data needs to be explored by tackling the data mining questions, which can be addressed using querying, reporting and visualization.
- Finally, the data quality must be examined by answering some important questions such as "Is the acquired data complete?", "Is there any missing values in the acquired data?"

Data preparation

The data preparation typically consumes about 90% of the time of the project. The outcome of the data preparation phase is the final data set. Once available data sources are identified, they need to be selected, cleaned, constructed and formatted into the desired form. The data exploration task at a greater depth may be carried during this phase to notice the patterns based on business understanding.

Modeling

- First, modeling techniques have to be selected to be used for the prepared dataset.
- Next, the test scenario must be generated to validate the quality and validity of the model.
- Then, one or more models are created by running the modeling tool on the prepared dataset.
- Finally, models need to be assessed carefully involving stakeholders to make sure that created models are met business initiatives.

Evaluation

In the evaluation phase, the model results must be evaluated in the context of business objectives in the first phase. In this phase, new business requirements may be raised due to the new patterns that has been discovered in the model results or from other factors. Gaining business understanding is an iterative process in data mining. The go or no-go decision must be made in this step to move to the deployment phase.

Deployment

The knowledge or information, which we gain through data mining process, needs to be presented in such a way that stakeholders can use it when they want it. Based on the business requirements, the deployment phase could be as simple as creating a report or as complex as a repeatable data mining process across the organization. In the deployment phase, the plans for deployment, maintenance and monitoring have to be created for implementation and also future supports. From the project point of view, the final report of the project needs to summary the project experiences and review the project to see what need to improved created learned lessons.

The CRISP-DM offers a uniform framework for experience documentation and guidelines. In addition, the CRISP-DM can apply in various industries with different types of data.

Data mining Architecture

Data mining is described as a process of discover or extracting interesting knowledge from large amounts of data stored in multiple data sources such as file systems, databases, data warehouses…etc. This knowledge contributes a lot of benefits to business strategies, scientific, medical research, governments and individual.

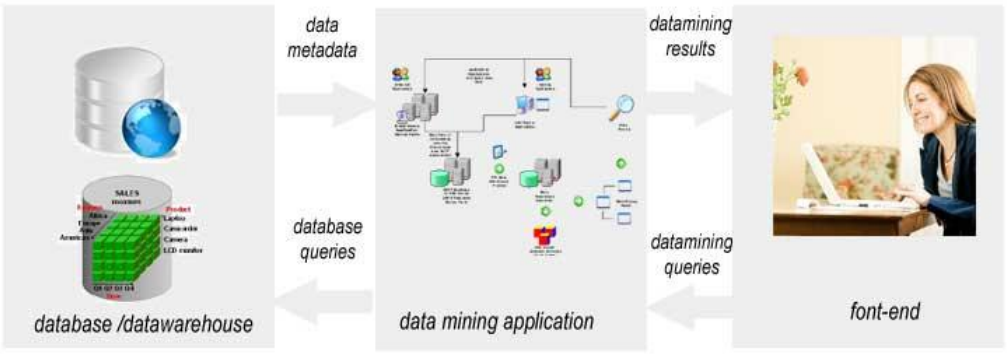
Business data is collected explosively every minute through business transactions and stored in relational database systems. In order to provide insight about the business processes, data warehouse systems have been built to provide analytical reports that help business users to make decisions.

Data is now stored in databases and/or data warehouse systems so should we design a data mining system that decouples or couples with databases and data warehouse systems? This question leads to four possible architectures of a data mining system as follows:

- **No-coupling**: in this architecture, data mining system does not utilize any functionality of a database or data warehouse system. A no-coupling data mining system retrieves data from a particular data sources such as file system, processes data using major data mining algorithms and stores results into file system. The no-coupling data mining architecture does not take any advantages of database or data warehouse that is already very efficient in organizing, storing, accessing and retrieving data. The no-coupling architecture is considered a poor architecture for data mining system however it is used for simple data mining processes.
- **Loose Coupling**: in this architecture, data mining system uses database or data warehouse for data retrieval. In loose coupling data mining architecture, data mining system retrieves data from database or data warehouse, processes data using data mining algorithms and stores the result in those systems. This architecture is mainly for memory-based data mining system that does not require high scalability and high performance.
- **Semi-tight Coupling**: in semi-tight coupling data mining architecture, beside linking to database or data warehouse system, data mining system uses several features of database or data warehouse systems to perform some data mining tasks including sorting, indexing, aggregation…etc. In this architecture, some intermediate result can be stored in database or data warehouse system for better performance.
- **Tight Coupling**: in tight coupling data mining architecture, database or data warehouse is treated as an information retrieval component of data mining system using integration. All the features of database or data warehouse are used to perform data mining tasks. This architecture provides system scalability, high performance and integrated information.

Let's examine the tight-coupling data mining architecture in a greater detail.

Tight-coupling data mining architecture



Data Mining Architecture

There are three tiers in the tight-coupling data mining architecture:

1. **Data layer**: as mentioned above, data layer can be database and/or data warehouse systems. This layer is an interface for all data sources. Data mining results are stored in data layer so it can be presented to end-user in form of reports or other kind of visualization.
2. **Data mining application layer** is used to retrieve data from database. Some transformation routine can be performed here to transform data into desired format. Then data is processed using various data mining algorithms.
3. **Front-end layer** provides intuitive and friendly user interface for end-user to interact with data mining system. Data mining result presented in visualization form to the user in the front-end layer.

Data Mining Techniques

There are several major ***data mining techniques*** have been developing and using in data mining projects recently including *association, classification, clustering, prediction, sequential patterns* and *decision tree*. We will briefly examine those data mining techniques in the following sections.

Association

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship between items in the same transaction. That's is the reason why association technique is also known as *relation technique*. The association technique is used in *market basket analysis* to identify a set of products that customers frequently purchase together.

Retailers are using association technique to research customer's buying habits. Based on historical sale data, retailers might find out that customers always buy crisps when they buy beers, and therefore they can put beers and crisps next to each other to save time for customer and increase sales.

Classification

Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups.

Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, we develop the software that can learn how to classify the data items into groups.

For example, we can apply classification in application that "given all records of employees who left the company, predict who will probably leave the company in a future period." In this case, we divide the records of employees into two groups that named "leave" and "stay". And then we can ask our data mining software to classify the employees into separate groups.

Clustering

Clustering is a data mining technique that makes meaningful or useful cluster of objects which have similar characteristics using automatic technique. The clustering technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes.

To make the concept clearer, we can take book management in library as an example. In a library, there is a wide range of books in various topics available. The challenge is how to keep those books in a way that readers can take several books in a particular topic without hassle. By

using clustering technique, we can keep books that have some kinds of similarities in one cluster or one shelf and label it with a meaningful name.

If readers want to grab books in that topic, they would only have to go to that shelf instead of looking for entire library.

Prediction

The prediction, as it name implied, is one of a data mining techniques that discovers relationship between independent variables and relationship between dependent and independent variables.

For instance, the prediction analysis technique can be used in sale to predict profit for the future if we consider sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction.

Sequential Patterns

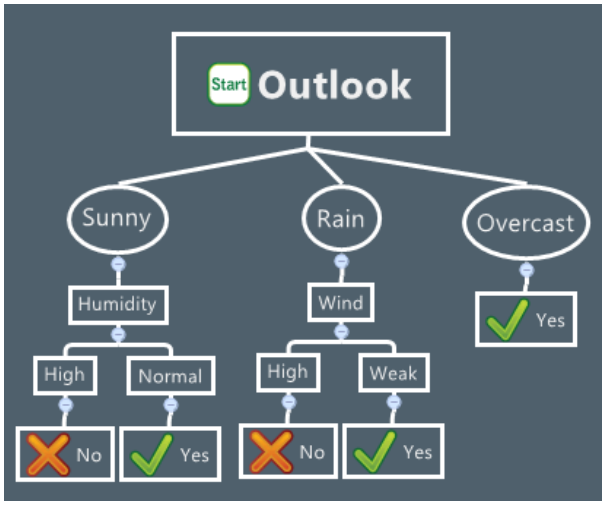
Sequential patterns analysis is one of data mining technique that seeks to discover or identify similar patterns, regular events or trends in transaction data over a business period.

In sales, with historical transaction data, businesses can identify a set of items that customers buy together a different times in a year. Then businesses can use this information to recommend customers buy it with better deals based on their purchasing frequency in the past.

Decision trees

Decision tree is one of the most used data mining techniques because its model is easy to understand for users. In decision tree technique, the root of the decision tree is a simple question or condition that has multiple answers.

Each answer then leads to a set of questions or conditions that help us determine the data so that we can make the final decision based on it. For example, We use the following decision tree to determine whether or not to play tennis:



Starting at the root node, if the outlook is overcast then we should definitely play tennis. If it is rainy, we should only play tennis if the wind is week. And if it is sunny then we should play tennis in case the humidity is normal.

We often combine two or more of those data mining techniques together to form an appropriate process that meets the business needs.

Application

Data mining is a process that analyzes a large amount of data to find new and hidden information that improves business efficiency. Various industries have been adopt data mining to their mission-critical business processes to gain competitive advantages and help business grows. This tutorial illustrates some data mining applications in sale/marketing, banking/finance, health care and insurance, transportation and medicine.

Data Mining Applications in Sales/Marketing

Data mining enables businesses to understand the hidden patterns inside historical purchasing transaction data, thus helping in planning and launching new marketing campaigns in prompt and cost effective way. The following illustrates several data mining applications in sale and marketing.

- Data mining is used for market basket analysis to provide information on what product combinations were purchased together, when they were bought and in what sequence. This information helps businesses promote their most profitable products and maximize the profit. In addition, it encourages customers to purchase related products that they may have been missed or overlooked.
- Retail companies uses data mining to identify customer's behavior buying patterns.

Data Mining Applications in Banking / Finance

- Several data mining techniques e.g., distributed data mining have been researched, modeled and developed to help credit card fraud detection.
- Data mining is used to identify customers loyalty by analyzing the data of customer's purchasing activities such as the data of frequency of purchase in a period of time, total monetary value of all purchases and when was the last purchase. After analyzing those dimensions, the relative measure is generated for each customer. The higher of the score, the more relative loyal the customer is.
- To help bank to retain credit card customers, data mining is applied. By analyzing the past data, data mining can help banks predict customers that likely to change their credit card affiliation so they can plan and launch different special offers to retain those customers.
- Credit card spending by customer groups can be identified by using data mining.
- The hidden correlation's between different financial indicators can be discovered by using data mining.
- From historical market data, data mining enables to identify stock trading rules.

Data Mining Applications in Health Care and Insurance

The growth of the insurance industry entirely depends on the ability of converting data into the knowledge, information or intelligence about customers, competitors and its markets. Data mining is applied in insurance industry lately but brought tremendous competitive advantages to

the companies who have implemented it successfully. The data mining applications in insurance industry are listed below:

- Data mining is applied in claims analysis such as identifying which medical procedures are claimed together.
- Data mining enables to forecasts which customers will potentially purchase new policies.
- Data mining allows insurance companies to detect risky customers' behavior patterns.
- Data mining helps detect fraudulent behavior.

Data Mining Applications in Transportation

- Data mining helps determine the distribution schedules among warehouses and outlets and analyze loading patterns.

Data Mining Applications in Medicine

- Data mining enables to characterize patient activities to see incoming office visits.
- Data mining helps identify the patterns of successful medical therapies for different illnesses.

Data mining applications are continuously developing in various industries to provide more hidden knowledge that increases business efficiency and grows businesses.

Advantages and Disadvantages of Data Mining

Data mining is an important part of knowledge discovery process that we can analyze an enormous set of data and get hidden and useful knowledge. Data mining is applied effectively not only in business environment but also in other fields such as weather forecast, medicine, transportation, healthcare, insurance, government…etc. Data mining has a lot of advantages when using in a specific industry. Besides those advantages, data mining also has its own disadvantages e.g., privacy, security and misuse of information. We will examine those **advantages and disadvantages of data mining** in different industries in a greater detail.

Advantages of Data Mining

Marketing / Retail

Data mining helps marketing companies build models based on historical data to predict who will respond to the new marketing campaigns such as direct mail, online marketing campaign…etc. Through the results, marketers will have appropriate approach to sell profitable products to targeted customers.

Data mining brings a lot of benefits to retail companies in the same way as marketing. Through market basket analysis, a store can have an appropriate production arrangement in a way that customers can buy frequent buying products together with pleasant. In addition, it also helps the retail companies offer certain discounts for particular products that will attract more customers.

Finance / Banking

Data mining gives financial institutions information about loan information and credit reporting. By building a model from historical customer's data, the bank and financial institution can determine good and bad loans. In addition, data mining helps banks detect fraudulent credit card transactions to protect credit card's owner.

Manufacturing

By applying data mining in operational engineering data, manufacturers can detect faulty equipments and determine optimal control parameters. For example semi-conductor manufacturers has a challenge that even the conditions of manufacturing environments at different wafer production plants are similar, the quality of wafer are lot the same and some for unknown reasons even has defects. Data mining has been applying to determine the ranges of control parameters that lead to the production of golden wafer. Then those optimal control parameters are used to manufacture wafers with desired quality.

Governments

Data mining helps government agency by digging and analyzing records of financial transaction to build patterns that can detect money laundering or criminal activities.

Disadvantages of data mining

Privacy Issues

The concerns about the personal privacy have been increasing enormously recently especially when internet is booming with social networks, e-commerce, forums, blogs…. Because of privacy issues, people are afraid of their personal information is collected and used in unethical way that potentially causing them a lot of troubles. Businesses collect information about their customers in many ways for understanding their purchasing behaviors trends. However businesses don't last forever, some days they may be acquired by other or gone. At this time the personal information they own probably is sold to other or leak.

Security issues

Security is a big issue. Businesses own information about their employees and customers including social security number, birthday, payroll and etc. However how properly this information is taken care is still in questions. There have been a lot of cases that hackers accessed and stole big data of customers from big corporation such as Ford Motor Credit Company, Sony… with so much personal and financial information available, the credit card stolen and identity theft become a big problem.

Misuse of information/inaccurate information

Information is collected through data mining intended for the ethical purposes can be misused. This information may be exploited by unethical people or businesses to take benefits of vulnerable people or discriminate against a group of people.

Conclusion

Comprehensive data warehouses that integrate operational data with customer, supplier, and market information have resulted in an explosion of information. Competition requires timely and sophisticated analysis on an integrated view of the data. However, there is a growing gap between more powerful storage and retrieval systems and the users' ability to effectively analyze and act on the information they contain.

Both relational and OLAP technologies have tremendous capabilities for navigating massive data warehouses, but brute force navigation of data is not enough. A new technological leap is needed to structure and prioritize information for specific end-user problems. The data mining tools can make this leap.

Quantifiable business benefits have been proven through the integration of data mining with current information systems, and new products are on the horizon that will bring this integration to an even wider audience of users.

Reference

- www.google.com
- www.wikipedia.com
- www.studymafia.org