ORIENTAL JOURNAL OF
COMPUTER SCIENCE & TECHNOLOGY
An International Open Free Access, Peer Reviewed Research Journal
Published By: **Oriental Scientific Publishing Co., India.**
www.computerscijournal.org

ISSN: 0974-6471
June 2012,
Vol. 5, No. (1):
Pgs. 69-73

An Overview of Data Mining

BASANTH KUMAR H.B.

Pooja Bhagavat Memorial Mahajana PG Centre, DoS in CS, KRS Road, Metagalli,
Mysore - 570 016 (India).

(Received: February 12, 2012; Accepted: June 04, 2012)

ABSTRACT

Organizations in the world wide generate huge amount of data which is mostly unorganized. This unorganized data requires some processing to generate meaningful and useful information. In order to organize the huge amount of data, we implement the database management system concept such as SQL Server. Structured Queury Language (SQL) is a query language used to retrieve and manipulate the data that are stored in relational database management systems. However, use of SQL is not always adequate to meet the end user requirements of sophisticated information from unorganized data bank. This paper describes the concepts of data mining, its process, techniques and some of its applications.

Keywords: Data Mining, KDD – Knowledge Data Discovery, patterns, machine learning.

INTRODUCTION

Data mining is a powerful technology with great potential to help researchers / organizations focus on the most important information in their large databases. Data mining tool predict future trends and behaviors, allowing business to make proactive, knowledge-driven decisions. The automated, prospective analysis offered by data mining move beyond the analysis 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.

Most of the organizations, who have massive quantities of data, started implementing data mining techniques rapidly on their existing software and hardware platforms to enhance the value of existing information resources and also to know how integration is possible with new products and systems as they are brought online. When implemented on high performance client/ server or parallel processing computers, data mining tools can analyze massive databases to predict the promising clients.

History of data mining

Data mining is fairly a new concept which was emerged in the late 1980s. but it soon attracted huge interests for research works and flourished with many new and remarkable techniques being discovered throughout the 1990s. Data mining in many ways is fundamentally the adaptation of machine learning techniques to business applications. Data mining is best described as the

union of historical and recent developments in statistics, Artificial intelligence and machine learning. These techniques are then used together to study the data and find the previously hidden trends or patterns within.

The evolution of database technology is as follows[1]

1950s: First computers, use of computers for census.
1960s: Data collection, database creation.
1970s: Relational data model.
1980s: Ubiquitous RDBMS, advanced data models and application oriented DBMS.
1990s: Data mining and data warehousing, massive media digitization, multimedia databases and web technology.

Data mining

Data mining is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data[2].

Knowledge Data Discovery (KDD)

Knowledge Discovery in Database (KDD) was formalized in 1989, with reference to the general concept of being broad and high level in the pursuit of seeking from data. The term data mining was then coined; this high-level application technique is used to present and analyze data for decision makers.

Fayyad et al. distinguish between KDD and data mining by giving following definitions. ***Knowledge Discovery in Databases*** is the process of identifying a valid, potentially useful and ultimately understandable structure in data. This process involves selecting or sampling data from a data warehouse, cleaning or preprocessing it, transforming or reducing it (if needed), applying a data mining component to produce a structure, and then evaluating the derived structure.

Data Mining is a step in the KDD process concerned with algorithmic means by which patterns or structures are enumerated from the data under acceptable computational efficiency limitations.

Thus, the structures that are outcome of the data mining process must meet certain conditions so that these can be considered as knowledge. These conditions are: validity, understandability, utility, novelty and interestingness[3].

Scope of data mining

Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

a) 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 very quickly.

b) Automated Discovery of previously unknown patterns

Data mining tools sweep through databases and identify previously hidden patterns[4].

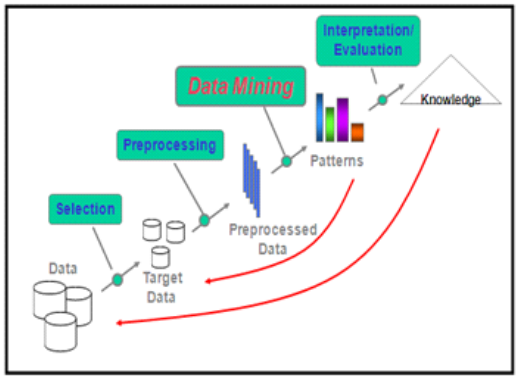

Fig 5. Process of KDD

Stages of Knowledge Data Discovery (KDD)

The stages of KDD, starting with the raw data and finishing with the extracted knowledge, are given below.

Selection

This stage is concerned with selecting or segmenting the data that are relevant to some criteria.

Preprocessing

In this stage the unnecessary information is removed.

Transformation

The data is not merely transferred across, but transformed in order to be suitable for the task of data mining. In this stage, the data is made usable and navigable.

Data Mining

This stage is concerned with the extraction of patterns from the data.

Interpretation and Evaluation

The patterns obtained in the data mining stages are converted into knowledge, which in turn, is used to support decision making. [2]

Data mining techniques

There are several major ***data mining techniques*** have been developed and used in data mining projects recently including association rules, classification, clustering, prediction and sequential patterns.

Association Rules

Association rules is a process to search relationship among data items in a given data set, which helps in managing all the data items. Association rules are used when a data set has a large data items. For example, a departmental store keeps a record of daily transactions where each transaction represents the items bought during one cash register transaction by a person. The manager of the departmental store generate the summarized report of the daily transactions that includes the information about what types of item sold at what quantity. This report also includes the information about those products, which are generally purchased together. Suppose the manager made a rule, if a person purchases the bread then he also purchase the butter. As a result, when availability of bread declines, probably the stock of butter also declines. The manager can make these types of checks using association rules[3].

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 make the software that can learn how to classify the data items into groups. For example, we can apply classification in application that "given all past records of employees who left the company, predict which current employees are probably to leave in the future." In this case, we divide the employee's records into two groups that are "leave" and "stay". And then we can ask our data mining software to classify the employees into each group[5].

Clustering

Clustering is a data mining technique that makes meaningful or useful cluster of objects that have similar characteristic using automatic technique. Different from classification, clustering technique also defines the classes and put objects in them, while in classification, objects are assigned into predefined classes. To make the concept clearer, we can take library as an example. In a library, books have a wide range of topics available. The challenge is how to keep those books in a way that readers can take several books in a specific topic without hassle. By using clustering technique, we can keep books that have some kind of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in a topic, he or she would only go to that shelf instead of looking the whole in the whole library[6].

Prediction

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

data mining technique that seeks to discover similar patterns in data transaction over a business period. The uncover patterns are used for further business analysis to recognize relationships among data.

Regression

Regression is a data mining technique used to fit an equation to a dataset. The simplest form of regression, linear regression, uses the formula of a straight line (y = mx + b) and determines the appropriate values for m and b to predict the values of y based upon given value of x. advanced techniques such as multiple regression, allow the use of more than one input variable and allow for fitting of more complex models, such as quadratic equation [7]. Regression uses existing values to forecast what other values will be[5].

Time series

Time series forecasting predicts unknown future values based on a time-varying series of predictors. Like regression it uses known results to guide its predictions [8]. Models must take into account the distinctive properties of time, especially the hierarchy of periods (including such varied definitions as the five or seven days work week, the twelveth – "month" year, etc), seasonality, calendar effects such as holidays, date arithmetic and special considerations such as much of the past is relevant.

Advantages of data mining

Marking/Retailing

Data mining can aid direct marketers by providing them with useful and accurate trends about their customers' purchasing behavior. Based on these trends, marketers can direct their marketing attentions to their customers with more precision. For example, marketers of a software company may advertise about their new software to consumers who have a lot of software purchasing history. In addition, data mining may also help marketers in predicting which products their customers may be interested in buying. Through this prediction, marketers can surprise their customers and make the customer's shopping experience becomes a pleasant one.

Retail stores can also benefit from data mining in similar ways. For example, through the trends provide by data mining, the store managers can arrange shelves, stock certain items, or provide a certain discount that will attract their customers.

Banking/Crediting

Data mining can assist financial organizations in areas such as credit reporting and loan information.

Law enforcement

Data mining can aid law enforcers in identifying criminal suspects as well as apprehending these criminals by examining trends in location, crime type, habit, and other patterns of behaviors.

Limitations of Data Mining

While data mining products can be very powerful tools, they are not self sufficient applications. To be successful, data mining requires skilled technical and analytical specialists who can structure the analysis and interpret the output that is created. Consequently, the limitations of data mining are primarily data or personnel related, rather than technology-related. Although data mining can help reveal patterns and relationships, it does not tell the user the value or significance of these patterns. These types of determinations must be made by the user. Similarly, the validity of the patterns discovered is dependent on how they compare to "real world" circumstances. For example, to assess the validity of a data mining application designed to identify potential terrorist suspects in a large pool of individuals, the user may test the model using data that includes information about known terrorists. However, while possibly re-affirming a particular profile, it does not necessarily mean that the application will identify a suspect whose behavior significantly deviates from the original model.

Another limitation of data mining is that while it can identify connections between behaviors and/or variables, it does not necessarily identify a causal relationship. For example, an application may identify that a pattern of behavior, such as the propensity to purchase airline tickets

just shortly before the flight is scheduled to depart, is related to characteristics such as income, level of education, and Internet use. However, that does not necessarily indicate that the ticket purchasing behavior is caused by one or more of these variables. In fact, the individual's behavior could be affected by some additional variable(s) such as occupation (the need to make trips on short notice), family status (a sick relative needing care), or a hobby (taking advantage of last minute discounts to visit new destinations)[9].

Applications

It is commonly used in a wide range of applications such as

1. **Telecommunication:** to handle transactional data such as phone call, data on mobile phones etc. other customer data such as billing, personal information and additional data such as network load, faults etc.
2. **Health:** to handle different aspects of the health system such as personal health records, hospital data and billing information.
3. **Astronomy:** to process terabytes of image and other data received from satellites.
4. **Economics and commerce:** analysis and prediction of stock market.
5. **Bioinformatics:** to predict diseases based on genome sequences.
6. **Terror, crime and fraud detections:** to find and predict the unusual events[4].

CONCLUSION

Wide-ranging 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. In this context, a new technological leap is needed to structure and prioritize information for specific end-user problems. The data mining tools can make this leap.

REFERENCES

1. U. Fayyad, Piatetsky-Shapiro, Smyth and Uthursamy (1996), Advances in Knowledge discovery and Data Mining, MIT press.
2. Data Mining techniques, Arun K Poojari, Universities Press, twenty first impression 2009.
3. Typical Data Mining Process for Predictive Modeling.
4. Data Mining: A Process to Discover Data Patterns and Relationships for Valid Predictions. volume no. 34, Issue No.9, December 2010.
5. Brieman, Friedman, Olshen and stone (1984), Classification and Regression trees, Wadsworth
6. Dorian Pyle (1999), Data Preparation for data Mining, Morgan Kaufmann.
7. Jiawei and Micheline kamber, Data Mining: Concepts and Techniques. Morgan Kaufmann publishers, 2001.
8. James Kobeilus (1 July 2008) the Forrester Wave: Predictive Analytics and Data Mining Solutions, Q1 2010, Forrester Research.
9. Jeffrey W. Seifert, Data Mining : An Overview, CRS Report for Congress, December 16, 2004.