DATA MINING: CONCEPTS AND TECHNIQUES IN DIGITAL AMBIANCE

P. S. Rajput * and Lala Ram Ahirwar**

* P. S. Rajput, Asstt. Librarian, Mohanlal Sukhadia University Udaipur(Raj.)

** Lala Ram Ahirwar, INDEST-AICTE Consortium, Central Library, I.I.T Delhi.

Abstract: *Data Mining, popularly known as Knowledge Discovery in databases is the automated or convenient extraction of patterns representing knowledge implicitly stored in large databases which solves the above problem. The study attempt to know what is data mining and why is it important. Deals about the basic concepts of data mining, its techniques and how work in digital atmosphere. Highlight the evolution of data mining and different types of step. Enumerates the problems associated with data mining in a digital ambiance.*

Keywords: Data mining, Knowledge discovery, Concepts, Techniques, and Digital Library.

Introduction

Data mining is currently regarded as the key element of a much more elaborated process called Knowledge Discovery in Databases (KDD). The knowledge is a collection of interesting and useful patterns in a database, whereas, KDD in a database is the non-trivial extraction of implicit, previously unknown and potentially useful information from data. The knowledge is stored in data warehouse, which is the central storehouse of data that has been extracted from operational data over a time in a separate database.

The information in a data warehouse is subject-oriented, non-volatile and of an historic nature, so they contain extremely large database. Data mining is the automatic extraction of patterns of information from these historical data.

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help library or information centers focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing 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 all queries that traditionally were too time consuming to resolve.

What is Data Mining?

Data Mining refers to extracting or “mining” knowledge from large amounts of data. Just like gold mining from rocks and sand, data mining should have been more appropriately named as “knowledge mining from data.” Since the term reflects emphasis on mining from large amounts of data, it is termed as “Data Mining.”

Data mining is the principle of sorting through large amounts of data and picking out relevant information. It is usually used by business intelligence organizations, and financial analysts, but it is increasingly used in the sciences to extract information from the enormous data sets generated by modern experimental and observational methods. It has been described as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data and "the science of extracting useful information from large data sets or databases".In short data mining means extraction of interesting information or patterns from data in large databases. Data mining is an essential step in the process of knowledge discovery in databases. The alternative names of data mining are Knowledge Discovery in Databases (KDD), Knowledge extraction, Data pattern analysis, Data archaeology, Data dredging, Information harvesting etc.
Data mining is a term used mainly in computer science. Sometimes it is also called knowledge discovery in databases (KDD). Data mining is about finding similarities in large sets of data. It is about discovering patterns in large sets of data. Very often, such data is stored in some form of database. Some commonly used algorithms can be classified as pattern-recognition, or Neural Network. Data mining is considered a subfield within the Computer Science field of knowledge discovery. Data mining is also closely related to applied statistics and its subfields descriptive statistics and inferential statistics

The term data mining is often used to apply to the two separate processes of knowledge discovery and prediction. Knowledge discovery provides explicit information that has a readable form and can be understood by a user. Forecasting, or predictive modeling provides predictions of future events and may be transparent and readable in some approaches (e.g. rule based systems) and opaque in others such as neural networks. Moreover, some data mining systems such as neural networks are inherently geared towards prediction and pattern recognition, rather than knowledge discovery

Evolution of Data Mining:

Data collection: Retrospective, status data delivery

Data Access: Retrospective dynamic data delivery at record level

Data ware housing: Retrospectives dynamic data delivery at multiple levels

Data mining: Prospective proactive information delivery

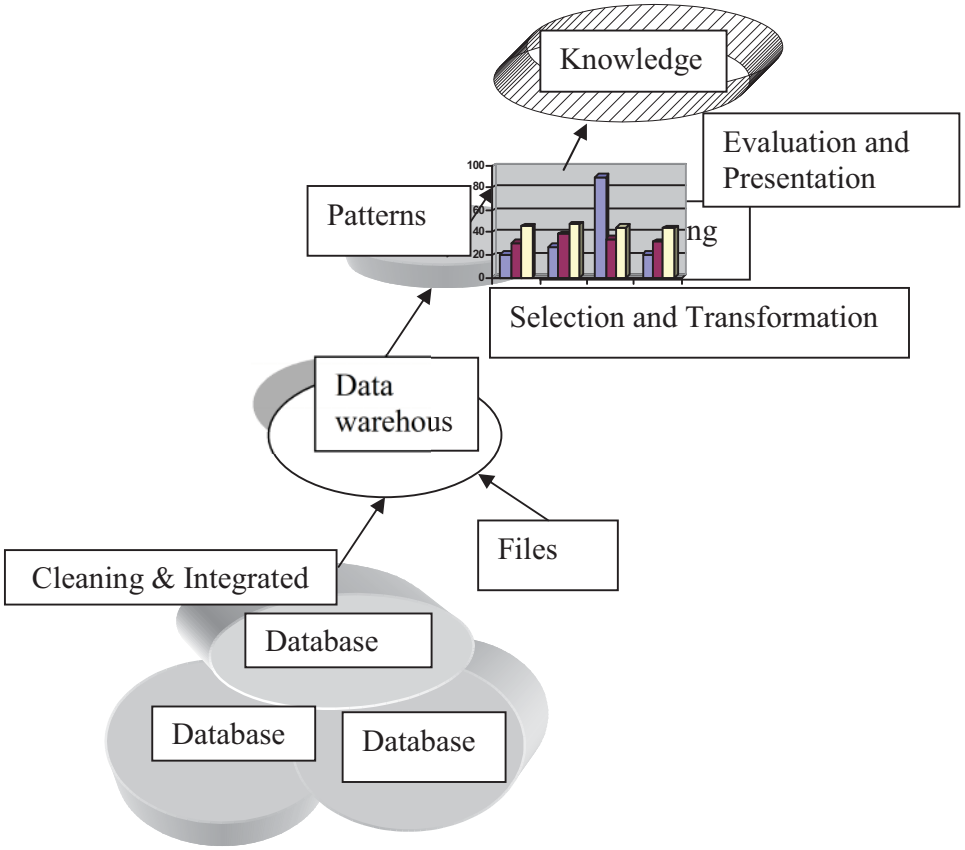
Importance

The major reason that data mining has attracted a great deal of attention in the information industry in recent years is due to the wide availability of huge amounts of data and the imminent need for running such data into useful applications ranging from business management, production control and market analysis to engineering design and science exploration. Data mining can be viewed as a result of the natural evolution of information technology. The steady and amazing progress of computer hardware technology in the past three decades has led to large supplies of powerful computers, data collection equipment, and storage media. Data can be stored in different types of databases, one of which is the data warehouse and it includes data cleaning, data integration and On-line Analytical Processing (OLAP). The fast growing tremendous amount of data, collected and stored in large and numerous databases, has far exceeded our human ability for Comprehension without powerful tools. Data mining is a powerful tool that overcomes all the above problems by extracting or mining knowledge from the large amount of data and provides the necessary information and knowledge to the users.
Data explosion problem evolved since automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories. We are drowning in huge data, but starving for knowledge. So data mining is the only solution to overcome the above problems, for extracting the huge data for getting useful information or patterns.

Steps in Data Mining Process

- Data gathering, e.g., data warehousing.
- Data cleansing: eliminate errors and/or inconsistent data
- Data selection; where data relevant to the analysis task are retrieved from the database.
- Data transformation;
- Data mining; an essential process where intelligent methods are applied in order to extract data
- Pattern evaluation; to identify the truly interesting patterns representing knowledge based on some interesting measures.
- Knowledge presentation; visualization techniques used to present the mined knowledge to the user.

Fig. 1: Data Mining Process Flow Chart



Technological Requirements

The basic infrastructure required for Data mining are: Mainframe system, client/serer and PC platforms enterprise-wide applications generally range in size form 10 gigabytes to over 11 terabytes. Capacity to deliver application exceeding 100 terabyte. There are two critical technological drivers.

- Size of the database: The more data being processed and maintained the more powerful the system required.
- Query complexity: the more complex queries and the greater number of queries being processed the more powerful system required.

Techniques Used in Data Mining:

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.

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

- **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). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

- **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*=1). Sometimes called the *k*-nearest neighbor technique.

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

- **Data visualization**: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

- 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 (On-Line Analytical Processing) platforms.

How DM Works

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought

- **Classes**: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.

- **Clusters**: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.

- **Associations**: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.

- **Sequential patterns**: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

Data Mining in Digital Ambience

Digital libraries of textural and multi-media data are now common and will soon be ubiquitous, while digital libraries of numerical data, especially tabular data are growing in importance. In this note, we discuss research challenges arising from the data mining of digital libraries. By data mining, we mean the automatic uncovering of patterns, associations, and anomalies in data.
As a simple motivating example, each quarter, public companies in the US file a 10-Q form with the Security and Exchange Commission (SEC). Most companies file this file electronically: it contains a variety of information about the financial health of the company, including tables containing the company's balance sheet, income statement, and cash flow statement. These forms are available on the web and can be retrieved by attribute (company name, company exchange symbol, date filed, etc.) or sometimes via full text search. Note that by retrieving documents in this way only refers to data contained within a single document. These may be thought as *intra-document* queries

Problems in Data Mining

There are some problems with data mining technology, they are:

- **Lack of Standards:** The most serious problem is that there are no established standards for data mining storage and retrieval. Beside, the record sharing between libraries is impractical and long term access to materials is in doubt. In an electronic environment where database access in a library is determined by IP or password validation, record sharing may not be an important consideration but information sharing programs such as interlibrary loan become complicated at beat and impossible at worst. Because data mining increase library dependency on proprietary functions, libraries that invest heavily in data mining technologies increase the risk of incurring expensive and difficult conversions or server data loss when vendors quit supporting their products. In the present environment, world wide use of the MARC format dramatically has reduced data migration problems and greatly simplified record sharing and interlibrary loan.
- **Unproven to Libraries:** It is unclear whether data mining techniques used on the Internet or for certain business and scientific applications can successfully apply in a library setting. In contrast to data mining in the business and scientific communities involve short documents consisting of well-structure or statistically oriented data, libraries work predominantly with large unstructured text document from diverse sources. While a number of text mining tools do provide access to minimally structured text document, the total amount of information they provide access to is small in comparison with that found in a large library.
- **Technical Hurdles:** The other problem with data mining is that it faces the same difficulties as other searching mechanisms. The quality of data is critical for successful data mining, just as it for successful searching by other method. If information is not structured in a way that allows pattern discovery, the likelihood of extracting meaningful information from the data is greatly reduced. Data mining looks for patterns in data. It is

very difficult for data mining tools to identify the relationship between different information objects when it is not possible to determine the meaning of the data. Despite of advancement in technology, it is not practical to use all processing techniques on all documents in a given search, except when small sets of data are concerned. Unless all data can be stored in memory and there is sufficient processing power, heuristics must be used to determine the optimal searching strategy. Users may revel information about themselves and to glean enough information to identify techniques that will optimally serve the user.

- **Inappropriateness of Data Mining Tools:** Before committing to data mining technologies on a large scale, libraries need to determine how data mining fits with existing resources and organizational goals. Generally speaking, data mining technologies are most beneficial to that are interested in purchasing access to databases rather that physical materials. Full-text, dynamically changing databases tends to be better suited to data mining technologies than the online catalog, which is cumbersome and expensive to update. On the other hand, libraries concerned with providing long-term access to physical items that exist within the library would be well advised to adopt a sit and wait attitude at this point especially since good access to these materials is provided through the online catalogue.

Some DM Systems

- Intelligent Miner: IBM data mining product
- Enterprise miner: SAS institute
- Mineset: Silicon graphics Inc.
- Clementine: Integral solution Ltd.
- DB miner: DB miner technology Inc.
- The data mining suite: Inf. Discovery Inc.

Conclusion

Generally, data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

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.

REFERENCES

Dhiman, Anil Kumar. Data Mining and its used in Libraries. *CALIBER*, Ahmedabad, February 13-15, 2003: 568-574.

Reilly, B. F. When Machines Do Research: Automated Analysis of News and Other Primary Source Text. *Journal of Library Administration,* 49, (5): July 2009, 507-517

Websites:

http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

http://en.wikipedia.org/wiki/Data_mining

http://www.thearling.com/text/dmwhite/dmwhite.htm

.
