ISSN No. 0976-5697

Volume 7, No. 6(Special Issue), November 2016

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Review of Data mining (Knowledge discovery) in the Future

Surender Kumar
Assistant Professor/Head (Computer Science)
S.G.T.B Khalsa College Sri Anandpur Sahib
Distt-Ropar (Punjab)
drsurender.sgtb@gmail.com

Kanwaldip Kaur
Assistant Professor
S.G.T.B Khalsa College Sri Anandpur Sahib
Distt-Ropar (Punjab)
kamal.wariech@gmail.com

Abstract-*Data mining (sometimes called data or knowledge discovery) 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.*

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 (e.g., association rule mining). 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. In Future different Scope of data mining are.

1. *Developing a unifying theory of data mining.*
2. *Scaling up for high dimensional data and high speed data streams.*
3. *Data mining in a network setting.*
4. *Data mining for biological and environmental problems.*
5. *Security, privacy and data integrity.*
6. *Dealing with non-static, unbalanced and cost-sensitive data*

DATA, INFORMATION, AND KNOWLEDGE:

DATA

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes**:**

- operational or transactional data such as, sales, cost, inventory, payroll, and accounting
- non operational data, such as industry sales, forecast data, and macro economic data
- meta data - data about the data itself, such as logical database design or data dictionary definitions

Information

The patterns, associations, or relationships among all this data can provide *information*. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

Knowledge

Information can be converted into *knowledge* about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

Data Warehouses

Dramatic advances in data capture, processing power, data transmission, and storage capabilities are enabling organizations to integrate their various databases into data *warehouses*. Data warehousing is defined as a process of centralized data management and retrieval. Data warehousing, like data mining, is a relatively new term although the concept itself has been around for years. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. Centralization of data is needed to maximize user access and analysis. Dramatic technological advances are making this vision a reality for many companies. And, equally dramatic advances in data analysis software are allowing users to access this data freely. The data analysis software is what supports data mining.

978-93-85670-72-5 © 2016 (RTCSIT)

CONFERENCE PAPER
International Conference on
Recent Trends in Computer Science & Information Technology (RTCSIT-2016)
21st August 2016
Guru Nanak College Budhlada, Punjab India

The knowledge discovery process:

Data mining is only one step of the knowledge discovery from databases (KDD) process. Data mining involves the application of techniques for distilling data into *information* or facts implied by the data. KDD is the higher level process of obtaining facts through data mining and distilling this information into knowledge or ideas and beliefs about the mini-world described by the data. This generally requires a human-level intelligence to guide the process and interpret the results based on pre-existing knowledge (Miller and Han 2001). The data mining is the critical interface between the syntactic knowledge or *patterns* generated by machines and the semantic knowledge required by humans for reasoning about the real world (Gahegan et al 2001).The KDD process does not seek any arbitrary pattern from a database; rather, data mining seeks only those that are *interesting*. These patterns are *valid* (a generalizable pattern, not simply a data anomaly), *novel* (unexpected), *useful* (relevant) and *understandable* (can be interpreted and distilled into knowledge) (Fayyad, Piatetsky-Shapiro and Smyth 1996). In addition to the scale of the data involved, the requirement for novelty distinguishes data mining from traditional statistics oriented towards hypothesis confirmation rather than generation. From a KDD perspective, anything that can be hypothesized *a priori* is not novel and therefore not interesting.

What can data mining do?

Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data. With data mining, a retailer could use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. By mining demographic data from comment or warranty cards, the retailer could develop products and promotions to appeal to specific customer segments.

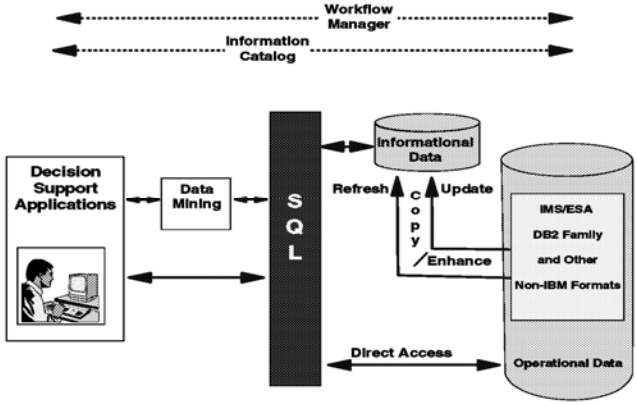
The National Basketball Association (NBA) is exploring a data mining application that can be used in conjunction with image recordings of basketball games. The Advanced Scout software analyzes the movements of players to help coaches orchestrate plays and strategies. By using the NBA universal clock, a coach can automatically bring up the video clips showing each of the jump shots attempted by Williams with Price on the floor, without needing to comb through hours of video footage. Those clips show a very successful pick-and-roll play in which Price draws the Knick's defense and then finds Williams for an open jump shot.

DATA MINING: VERIFICATION VS. DISCOVERY:

Decision support systems (DSS), executive information systems, and query/report writing tools are used to produce reports about data, usually aggregating it through any number of dimensions. Another use of these tools is to detect trends and patterns in customer data that will help answer some questions about the business. When used in this mode, a query is created to access the records relevant to the question(s) being formulated. After the data is retrieved, it is examined to detect the existence of patterns or other useful information that can be used in answering the original question(s). We call this *the verification mode.* In this mode, the user of a DSS generates an hypothesis about the data, issues a query against the data and examines the results of the query looking for affirmation or negation of the hypothesis. In the first case, the process ends; in the latter case, a new query is reformulated and the process iterates until the resulting data either verifies the hypothesis or the user decides that the hypothesis is not valid for his data.

Data Mining and the Information Warehouse framework

Data mining tools discover useful facts buried in the raw data (thus the term discovery model.) They complement the use of queries, multidimensional analysis and visualization tools to gain a better understanding about data. As such, good facilities to perform queries and data visualization as well as the availability of powerful data mining operators should be part of a well architected Decision Support environment. shows such an architecture. Much like a regular mining process, which takes raw material as it may exist in a mine and through several steps extracts from the ore valuable metals, data mining comprises three distinct phases or steps< Data Preparation, Mining Operations and Presentation> The process of information discovery can be described as an iteration over the three phases of this process.



DATA MINING IN AN INFORMATION WAREHOUSE ENVIRONMENT

The first phase, Data Preparation, can be further split into two: Data Integration and Data Selection and Pre-analysis. Data Integration refers to the process of merging data which typically resides in an operational environment having multiple files or databases. Resolving semantic ambiguities, handling missing values in data and cleaning dirty data sets are typical data integration issues. Because these issues are common with those found while building Data Warehouses, we will not discuss them here. A discussion of these topics is found in the companion white paper "IBM Information Warehouse Solution: A Data Warehouse - PLUS." Data Mining does not require that a Data Warehouse be built. Often, data can be downloaded from the operational files to flat files that contain the data ready for the Data Mining analysis. However, in many situations, like the one shown in , Data Mining can and will be performed directly from a Data Warehouse. Other issues that occur during integration that are specific to Data Mining deal with identifying the data required for mining and eliminating bias in the data. Identifying the data that is relevant to a given mining operation is a problem for which there is no good solution in the marketplace. The person doing the analysis has to determine which data is relevant to the mining operation being performed. For example, to discover product affinities in market basket analysis one may include information about advertising and shelf placement. Bias in the data can result in the "discovery" of erroneous information. For this reason bias in data should be detected and removed prior to performing the mining operations.

Visualization and knowledge discovery

Visualization is a powerful strategy for leveraging the visual orientation of sighted human beings. Sighted humans are extraordinarily good at recognizing visual patterns, trends and anomalies; these skills are valuable at all stages of the knowledge discovery. Visualization can be used in conjunction with OLAP to aid the user's synoptic sense of the database. Visualization can also be used to support data preprocessing, the selection of data mining tasks and techniques, interpretation and integration with existing knowledge (Keim and Kriegel 1994). Visualization creates an opportunity for machines and humans to cooperate in ways that exploit the best abilities of both (fast but dumb calculation and record-keeping versus slow but smart recognition and interpretation, respectively) (Gahegan et al. 2001).

How does data mining work?

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 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.

Different levels of analysis are available:

- **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.

978-93-85670-72-5 © 2016 (RTCSIT)

- **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.

What technological infrastructure is required?

Today, data mining applications are available on all size systems for mainframe, client/server, and PC platforms. System prices range from several thousand dollars for the smallest applications up to $1 million a terabyte for the largest. Enterprise-wide applications generally range in size from 10 gigabytes to over 11 terabytes. NCR has the capacity to deliver applications exceeding 100 terabytes.

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 the queries and the greater the number of queries being processed, the more powerful the system required. Relational database storage and management technology is adequate for many data mining applications less than 50 gigabytes. However, this infrastructure needs to be significantly enhanced to support larger applications. Some vendors have added extensive indexing capabilities to improve query performance. Others use new hardware architectures such as Massively Parallel Processors (MPP) to achieve order-of-magnitude improvements in query time. For example, MPP systems from NCR link hundreds of high-speed Pentium processors to achieve performance levels exceeding those of the largest supercomputers.

CONCLUSION

Data mining is not a new phenomenon. All large organizations already have data mining, but they are just not managing them. It is very helpful as it provides the means to change raw data into information for making effective decisions. Data in the data mining is preprocessed and presented such that it facilitates the cross functional monitoring and assessment of the overall direction of the organization. Thus, it is the hub for an intelligent management decision support.

Successful implementation of a data mining requires a high-performance, scaleable combination of hardware and software, which can integrate easily with existing systems, so that users can use data mining to improve their decision-making.

A data mining is incomplete until it provides the exploitation tools that enable end users to view, analyze and report on data in ways that support their decision-making. Data marts, data modeling and metadata are some other important concepts attached with data mining , the knowledge of which helps to a great extent in data mining implementation.

REFERENCES

[1] Frawley and G. Piatetsky-Shapiro and C. Matheus (Fall 1992). "Knowledge Discovery in Databases.

[2] Hand, H. Mannila, P. Smyth (2001). *Principles of* Data Mining.

[3] “From data mining to knowledge discovery: An overview” in U.M. Fayyad, Piatetsky-Shapiro, P. Smyth and R. Ulthurusamy (eds.) (1996).