Journal of Basic and Applied Engineering Research
p-ISSN: 2350-0077; e-ISSN: 2350-0255; Volume 5, Issue 3; January-March, 2018, pp. 155-158

http://www.krishisanskriti.org/Publication.html

Data Mining Technology

S. Shahabuddin

University Department of Physics & Computer Science
Veer Kunwar Singh University, Ara

Abstract—*Data mining on large databases has been a major concern in research community, due to the difficulty of analyzing huge volumes of data using only traditional OLAP tools. Mining information and knowledge from large database has been recognized by many researches as a key research topic in database system and machine learning and by many industrial companies as an important area with an opportunity of major revenues. Researchers in many different fields have shown great interest in data mining. Several emerging application in information providing services, such as data warehousing and online services over the Internet, also call for various data mining techniques to better understand user behavior, to improve the service provided and to increase the business opportunities. This sort of process implies a lot of computational power, memory and disk I/O, which can only be provided by parallel computers. We present a discussion of how database technology can be integrated to data mining techniques.*
Data mining is a process consisting in collecting knowledge from databases or data warehouses and the information collected that had never been known before, it is valid and operational. Nowadays data mining is a modern and powerful IT&C tool, automatizing the process of discovering relationships and combinations in raw data and using the results in an automatic decision support.

Keywords*: data mining, data warehouse, knowledge discovery, OLAM, OLAP.*

1. INTRODUCTION

Data mining technique have increasingly been studied especially in their application in real-world databases one typical problem is that database stand to be very large and these techniques often repeatedly scan the entire set sampling has been used for a long time but subtle

Differences among sets of objects become less evident. Knowledge discovery and data mining have emerged as an interdisciplinary domain with a fast evolution and merging with databases, statistics, data warehouses and willing to extract a big amount of valuable knowledge and information.

The purpose of data mining is to discover unknown new information and due to this fact, the results are truly useful. The knowledge discovered through data mining must be valid. Appling the data mining techniques on large amounts of varied data could lead also to false information therefore is essential to check the data validity. We could also refer to the data mining process as a step in discovering the information through a set of algorithms and patterns meaningful in the data structures and showing market trends.

Data mining discovers patterns within data, using predictive techniques. These patterns play a very important role in the decision making because they emphasize areas where business processes require improvement. Using the data mining solutions, organizations can increase their profitability, can detect fraud, or may enhance the risk management activities. The models discovered by using data mining solutions are helping organizations to make better decisions in a shorter amount of time.

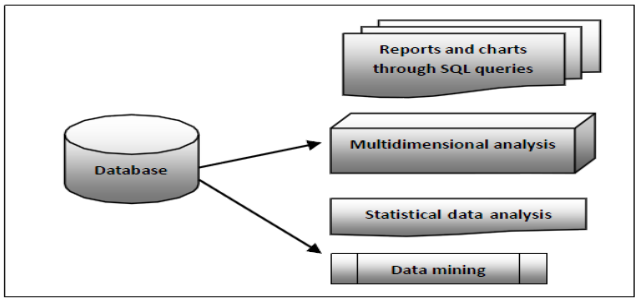

Figure 1 - Differences between traditional data analysis and data mining

Data mining methods derived from statistical calculation, database administration and artificial intelligence, SQL queries, analysis in multidimensional databases using OLAP systems. They do not replace the traditional methods of statistics, but are considered to be extensions of graphic and statistical techniques.

Typical data structure suitable for data mining contains the observations placed on lines and the variables placed on columns. Domains or range values for each variable must be precisely defined, avoiding as much as possible vague expressions. Line and column format, similar to the spreadsheet file is required for data mining.

Data mining software is separated into two groups:

- Data mining tools – are providing techniques that can be applied to any business problems.

- Data mining applications – incorporate techniques inside an application specially built to address business problems. Our life is influenced by data mining applications. For example, almost any financial transaction is processed by a data mining application to detect fraud. Increasingly more organizations are using both data mining tools and applications to develop predictive analysis.

2. OVERVIEW OF DATA MINING TECHNIQUE

Data mining is a step in knowledge discovery in databases (KDD) that searches for a series of hidden patterns in data often involving a repeated iterative application of particular data mining methods the goal of the whole KDD process is to make patterns understand able to humans in order to facilitate a better interpretation of the underlying data

We present four classes of data mining techniques typically used in a variety of well-known applications and researches currently cited in the database mining community they certainly do not represent all mining methods but are a considerable portion of them when a large amount of data is considered.

2.1. Classification

Classification is a well-known data mining operation and it has been studied in machine learning community for long time. Its aim is to classify cases into different classes based on common properties (attributes) among a set of objects in a database. After the construction of the classification model, it is used to predict classes of new cases that are going to be inserted in the database adequate applications for classification include medical diagnosis credit risk assessment fraud detection and target marketing

2.1.1 Decision Trees

Decision tree methods area kind of machine learning algorithm that uses a divide-and-conquer approach to classify cases using a tree based representation. They usually use a greedy algorithm that recursively sub divides the training set until reaching a partition that represents cases totally belonging to the same class or until a criteria is reached(pre-pruning) .When deciding what attribute is going to be used by each subdivision a statistical test is adopted as the splitting criteria

2.1.2 Neural Network

Method based on artificial network provides a general and practical method for learning functions. Which is represented by continuous attributes discrete or vectors One important characteristic of the algorithm is its robustness when dealing with errors in the training set .Basically neural networks have been used to interpret visual scenes voice recognition and they are not only used for classification(e.g. neural networks are widely used for prediction purposes).

2.2 Association Rules

Mining association rules is particularly important when trying to find relevant associations among items in a given customer transaction An example of the output of such mining is the statement that 80% of transactions that purchase diapers and milk also purchase milk bottles The number 80%istheruleconfidencefactor

2.3 Clustering

Clustering algorithms also called unsupervised classification is the process of grouping physical or abstract objects into classes of similar objects Clustering analysis helps to construct meaningful partitioning of a large set of objects based on a divide and conquer methodology, which decomposes a large-scale system into smaller components to simplify design and implementation.

2.4 Sequential Patterns

The discovery of sequential patterns has been motivated by applications in retailing industry including attached mailing and add-on sales and in the medical domain.

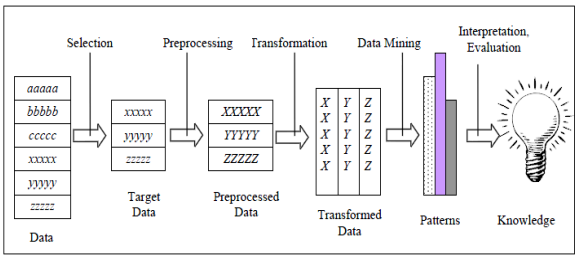

Figure 2: The Data Mining Process

3. DATA MINING AND KNOWLEDGE DISCOVERY COMPONENTS

The main function of data mining is to extract knowledge patterns from data. Therefore, data mining uses a variety of statistic algorithms, forms recognitions, classifications, fuzzy logic, machine learning, genetic algorithms, and neural networks. The variety of algorithms can be grouped into the main components of data mining.

The main components of data mining are:

1. The model, which, like any other computerized model, is represented by a function in single-dimensional or multidimensional space, depending on the parameters. It may be represented either as a linear function of parameters either as a probability or fuzzy function. Different algorithms, such as classification and clustering are leading to the achievement of the model.
2. Preference criteria may have a different nature, some of them being based on ranking and others on interpolation or on the best approximation.

3. Selection algorithms are leading to the selection of three important elements that occur in data mining: the model selected from a model base, data selected from the database and setting up the parameters and the preference criteria, selected from a criteria base.
4. Setting residuals generally consist in algorithms of determination of deviation and stability; a particular category of such algorithms is the statistical ones, which are setting the deviations from the ideal model.

Each commercial product uses several algorithms and in each of them, we can find some or all of the above components, in different proportions.

Researches who make the difference between data mining and knowledge discovery consider knowledge discovery as a complex interactive and iterative process, which includes data mining. Thus, it is considered that the knowledge discovery retrieval is accomplished in the following steps:

a) Understanding the applicability domain and the formulation of the problem. This step is an essential condition in extracting relevant knowledge for choosing the most suitable method of data mining for the third stage, according to the destination of the application and the nature of data.
b) Collecting and reprocessing data, including the selection of data sources, removing the outer layers, processing and data reduction.
c) Step third is represented by data mining, the process of extracting models or patterns hidden in data. A model is a global representation of a structure that summarizes the systematic component, underlying the data, or which describes how data can result. A pattern is a local structure, associated with some variables and conditions. The most important data mining methods are classification and predictive regression modeling, clustering, dependences modeling with graphical models and estimation of density.
d) The fourth step is the interpretation or post processing of knowledge found, especially the interpretation in terms of description and prediction, the two main purposes in the discovery system practice. Experience shows that patterns or data patterns are not directly used and the knowledge discovery process is repeated through the knowledge discovered. A standard manner of assessment is to divide data into two sets, working on a data set and testing on the second. We repeat the process a number of times, each time dividing the data differently. The results will be used to estimate the rules of performance.
e) The final step is to put into practice the knowledge discovered. In some cases, the new discovered data can be used without the need of an integrated system and, in other cases; it can be used to exploit it through specialized software.

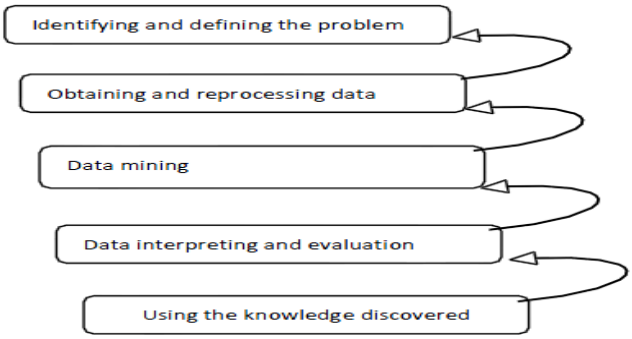

Fig. 3: Knowledge extraction process

4. ON-LINE ANALYTICAL DATA MINING SYSTEMS (OLAM)

Data mining and OLAP system are tools for business intelligence. OLAP queries retrieve the database information, at certain levels. OLAP analysis is a deductive process. Based on this hypothesis, data mining is different from OLAP system because it is using its data to discover new patterns. This tool examines the data and interactions between them.

Data mining technology is focused on assessing the predictive power of patterns, this being possible by testing conclusions on a different set of data and by calculating the predictive accuracy. Data mining could help analyse and design the data warehouse by focusing attention on important variables, identifying exceptions and finding interactions between variables. Due to the interconnection between the two technologies, the OLAM systems have emerged. OLAM systems are also called OLAP systems for data mining. This type of system integrates OLAP multidimensional processing with extracting knowledge form data, in data mining.

Lately, different architectures had been defined but OLAP systems are imposed increasingly due to their advantages:

- Ensuring a high quality of data in the data warehouse.
- Exploiting the data processing infrastructure, available in the data warehouses. Processing facilities provided by the data ware house includes accessing, integration, consolidation and transformation of the heterogeneous databases, multiple Web access, report generating and online analysis
- Analyzing data based on the facilities offered by OLAP processing.
- Selecting online the processing functions for the data mining.

An OLAM system must retrieve knowledge in multidimensional data in the same way OLAP systems carries out the data processing.

5. CONCLUSION

Data mining initially generated a great deal of excitement and press coverage, and, as is common with new “technologies”, overblown expectations. However, as data mining has begun to mature as a discipline, its methods and techniques have not only proven to be useful, but have begun to be accepted by the wider community of data analysts

Data mining tools and applications are helpful in business management, business intelligence, selective marketing, and decision analysis.

Data mining is a technology that uses complex and elaborate algorithms in order to analyse and reveal interesting information useful in the analysis made by decision makers. OLAP organizes data into a pattern suitable for the analysts to operate while data-mining carries out data analyses and provides the results to the decision makers. Thus, OLAP enables a model-oriented analysis while data mining makes the oriented data analysis easier.

REFERENCES

[1]Berry, M., and Linoff, G. S. *Data Mining Techniques for Marketing, Sales, and Customer Relationship Management.* Wiley. (2004).

[2]Han, Jiawei and Kamber, Micheline. *Datamining: concepts and techniques.* : Morgan Kaufmann Publishers, 2006.

[3 [Bigus.J.P...*Data Mining with Neural Networks* .McGraw-Hill. 1996.

[4] Lukasz Kurgan and Petr Musilek.*A survey of Knowledge Discovery and Data Mining process models* (2006).

[5] Gorunescu, Florin. *Data Mining Concepts, Models and Techniques Springer*, 2011.

[6] Chakrabarti, S... *Mining the Web: Statistical Analysis of Hypertext and Semi-Structured Data.* Morgan Kaufmann (2002).

[7] Witeen, Ian H.; Frank, Eibe; Hall Mark A. *Data Mining: Practical Machine Learning Tools & Techniques* (Jan 2011).

[8] C. C. Aggarwal, *Data Mining, Cham: Springer International Publishing*, 2015.

[9] G. B. Achary, P. Venkateswarlu, B. V. Srikanth, *"Importance of HACE and Hadoop among Big data Applications", Int. J. Res.*, vol. 2, no. 3, pp. 266-272, Mar. 2015.

[10] G. K. Gupta, *Introduction to data mining with case studies*, PHI Learning Pvt. Ltd., 2014.

[11] Y. Zhao, H. Lin, "*WEB data mining applications in e-commerce*", *9th International Conference on Computer Science and Education*, pp. 557-559, 2014

https://www.invensis.net/blog/data-processing/12-data-mining-tools-techniques/