International Journal of Advanced Research in IT and Engineering

ISSN: 2278-6244

INTRODUCTION TO DATA MINING

Hedayat Bahadori*

Sara Bahadori*

Shabnam Azari*

Abstract: *The subject of data and information has been considered by organizations and managers for a long time and experts attempt to discover and extract the hidden knowledge in this data. Nowadays, organizations are successful that they access to this information easily and correctly. Correct management of obtained knowledge, analyzing inside and outside the organization and obtaining the information and discovering knowledge from this data and transform it into organizational knowledge will be factors for obtaining competitive benefits and improving organizational process and finally promoting the quality of products/services for superiority over the competition. This paper examines the importance and necessity of data mining.*

*Department of Computer Engineering, Omidiyeh Branch, Islamic Azad University, Omidiyeh, Iran

1. INTRODUCTION

Technical abilities of human beings have been increased quickly for manufacturing and collecting data in two recent decades. Factors such as the widespread use of barcode for commercial manufacturing, employing computer in business, sciences, governmental services and improving tools and collecting data play significant roles in these changes [1].

Public use of web and internet as a global information supply systems lead human beings to encounter a large volume of data and information. Explosion growth in stored data has created an urgent need to new technologies and automatic tools helping human beings intelligently in order to transform this large volume of data into information and knowledge. Data mining has been raised as a resolution for these problems. Data mining enjoys several scientific fields simultaneously. For instance, we can mention technology of database, artificial intelligence, machine learning, neutral network, statistic, recognition of pattern, knowledge- based system, knowledge acquisition, information retrieving, high performance computing, data visualization [2].

Data mining is a step of process of knowledge discovery and includes special algorithms of data mining how discovers patterns or models in data under effective constraints of plausible computation [3]. Simply, data mining refers to extraction process of unknown, useful and potential knowledge of data.

2. HISTORY

Regarding to available precious information in database in the late 1980s, extracting and use of information of database were started. Data mining is a process that appeared at the beginning of 1990s and it deals with extracting information from database by new theory. Discovering knowledge workshops were held in 1989 and 1991 by Pia Ttesky and his colleagues and mentioned workshops were held by Fayyd & Pia Ttesky between 1991 & 1994. The term "data mining" has officially been raised for the first time by Fayyad in the first international conference of "Discovering knowledge and data mining" in 1995. Data mining has purposefully been brought in statistic subject since 1995 and the first issue of discovering knowledge was published from database. Nowadays, several conferences are held through the world in this field. Data mining has historically been gradual evolution and database has been raised as a science since the early 1990s simultaneously using database over the world.

3. WHAT IS THE SUBJECT OF DATA MINING?

The subject of data mining is to know new and precious, potential, useful, logical relations things and available patterns in data. Finding useful patterns in data with several titles (like data mining) are stated in several societies. For example, titles such as extracting knowledge, discovering information, removing information, processing patterns of data can be mentioned.

The term of data mining is used by statisticians, researchers of databases and management information systems and commercial associations. Data mining is resulted in traditional analysis of data and statistical approaches how includes analytical techniques creating other branches, like:

- Numerical analysis
- Consistence and surfaces patterns from artificial intelligence like machine learning
- Neutral networks and genetic algorithms etc.

However, many data mining emphasizes on traditional methods and approaches of data analysis based on theory. There are basically two approaches for data mining that they differ from each other regarding to creating and designing model and finding patterns. The first approach that is related to manufacturing model (separating problems exist in great databanks) is the same as common statistical investigative method. In this method, the goal is that we obtain total summaries from databanks for recognition and description of main properties of distribution form. There are examples such models that include analysis of sectional cluster from databanks of regression model for prediction and rule of ranging with tree structures.

Second type of approach of data mining is approach of diagnosis of pattern. This approach has attempted to diagnose (which any case they are important) offsets although small one (than optimal) to disappear unusual process and pattern. Examples such as unusual patterns (to detect fraud) using credit cards and subjects that patterns are the same as non-identical characteristics with other patters are kinds of these applications. Such strategies that lead data mining to be considered as science of precious information search among large mass of data. Generally, in business (commercial) databases, the weakness of understanding patterns are due to their complexities. These complexities are created due to discontinuous, unclear, incompletion. Although, most of data mining algorithms can distinguish the effect

of irrelevant properties in diagnosis of main pattern. The power of prediction of data mining algorithms decrease by increasing these offsets.

4. WHAT DOES CAUSE DATA MINING TO BE CREATED?

The most important considerable reason of data mining in information industry is to access high volume of data and need to extract purposive information and knowledge from this data. Obtained information and knowledge are used in enormous applications from business management and manufacturing control and market analysis to engineering designing and scientific researches. Data mining is resulted in natural evolution of information technology. This evolution leads to improve in database industry like operation of collecting data and establish database, manage data and analyze and understand data [4].

Evolution of data mining technology and frequency of use of it in different application cause to collect high volume of data. Frequency data cause to need to powerful tools for analyzing data because currently human beings do not have enough information regarding to well data. Data mining tools analyze data and discover data templates. We can use its results in applications like determining strategy for business, knowledge base and scientific and medical researches. Available gap between data and information cause data mining tools to need in order to convert worthless data to worth knowledge (Hand & Manila, 2001). Data mining simply means the extraction or (mining) of knowledge from many raw data. Of course, this name is not mostly appropriate. Because, for example operation of mining for extracting gold from rock and sand is called gold mining, neither sand mining nor rock mining. Therefore, it is better to be called this process “extracting knowledge from data” that is unfortunately so long. “Knowledge mining” as shorter term as substitution can not explain emphasis and importance on mining of many volume of data. The term mining leads human beings to remember the process of finding small set of precious parts from high volume of raw materials (Hand & Manila, 2001).

5. STEPS OF KNOWLEDGE DISCOVERY

Knowledge discovery includes following frequentative steps:

1- Data cleaning (removing noise and inconsistencies of data)

2- Data integration (several data source are combined)

3- Data selection (data related to analysis are retrieved database)

4- Data transformation (transforming data into a form that is appropriate to data mining like summary and aggregation)
5- Data Mining (main process that apply intelligent processes for extracting templates from data)
6- Pattern evaluation (for determining correct and foresaid patterns by measurement criterion)
7- Knowledge presentation (for visual display, techniques of revealing knowledge is used in order to present discovered knowledge to user).

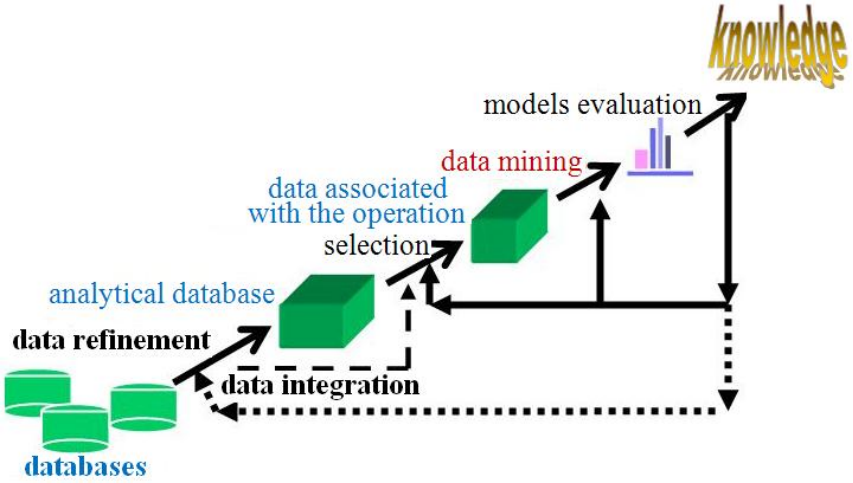

Figure 1: Steps of operation of knowledge discovery from database

Each data mining steps should interact to user or knowledge base, discovered patterns are presented to user and it is added as knowledge to knowledge base regarding his/her request. According to this attitude data mining is only step from total processes. Of course as a basic step that reveals the hidden patterns. According to presented subjects, we give here the definition of data mining:

"Data mining includes the process of finding knowledge from high volume of stored data in database, data repository or other information reservoirs" [4]. Based on this attitude, a data mining system typically includes following main components. Figure 2 represents architecture of system.

1- Database, data reservoir or other information reservoirs: are sets of databases, data reservoirs, spreadsheet or other information reservoirs that data cleaning, integration techniques are done on this data.

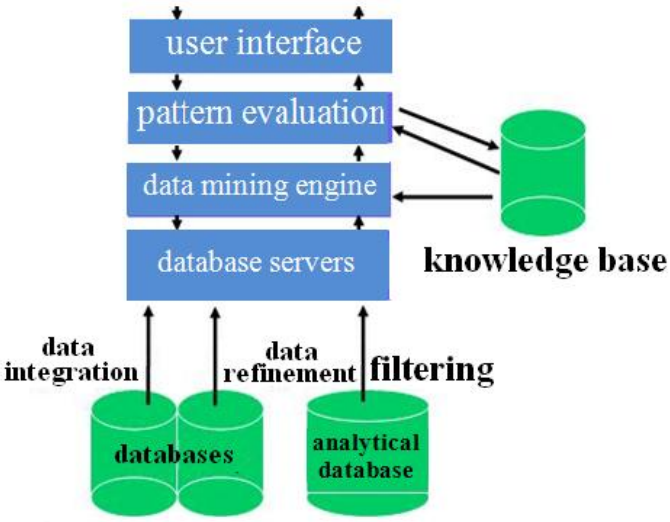

Figure 2: Architecture of a data mining system

2- Databases or data reservoir servers: are responsible for retrieving data related to type of user request of data mining of user.

3- Knowledge base: this base contains of domain knowledge in order to help searching and it is also used for retrieving found patterns.

4- Data mining engine: this engine is the main part of data mining system. It ideally includes modules like characterization, association, classification, cluster analysis, evolution and deviation analysis.

5- Pattern evolution module: applies interesting measures criterion and interact with module of data mining how it focuses on searching among interesting patterns. Using a threshold interesting to evaluate discovered patterns.

6- Graphical user interface (Gui): This module communicate between user and data mining system. It allows the user to communicate with data mining system by query.

5. CONCLUSION

Doing data mining process, high information or knowledge is extracted from database and it will be investigable from various attitudes. Discovered knowledge in decision systems, process control, information management and query processing will be useful [4].

Therefore, data mining is considered as one of pioneer branches in information industry and is one of the most promising fields of developing in information industry.

REFERENCES:

1. hisao ishibuchi, "*Introduction to Data Mining and Knowledge Discovery*", Two Crows Corporation. 1999.
2. David Hand, Heikki Mannila , Padhraic Smyth, "*Principles of Data Mining*". The MIT Press, 2001.
3. Tomoharu nakashima and hisao ishibuchi. 2005 : "*Using Boosting Techniques To Improve The Performance Of Fuzzy Classification Systems*" In "*Classification And Clustering For Knowledge Discovery*". Studies In Computational Intelligence. Vol 4. pp.146-157 . Springer, Netserlands.
4. J.Han, and M.Kamber. 2001 : "*Data Mining: Concepts and Techniques*". San Diego Academic Press.