International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-3, Issue-2)

Research Article | **February 2014**

Data Mining: A Literature Survey

Meenakshi Sharma
Global Research Institute of Management.& Technology,
Radaur, Haryana, India

Abstract*— Data mining is a powerful new technology to discover information within the large amount of the data. Data mining is considered as an important subfield in knowledge management. Today, Data mining helps different organization focus on the information in the data they have collected about the behaviour of their customer's. From last few years, research in data mining continues growing in various fields of organization such as Statistics, Machine Learning, Artificial Intelligence, Pattern Recognition, business, education, medical, scientific etc. In this paper, discusses the concept of data mining, important issues and applications.

***Keywords*— Data mining, Data Base, KDD, AI, Information**

I. INTRODUCTION

Data mining [1] is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques."Data mining sometimes called data or knowledge discovery. 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.

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information the patterns, associations, or relationships among all this *data* can provide *information.*

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. As a advancement of information technology in various fields of human life has increased to the large amount of data storage in various ways like records, documents, images, sound recordings, videos, scientific data, and many new data formats. For better decision making, the large repositories data collected from different resources require proper mechanism of extracting knowledge from the databases. Knowledge discovery in databases (KDD) [2], often called data mining, extracting information and patterns from data in large data base. The core functionalities of data mining are applying various techniques to identify nuggets of information of decision making knowledge in bodies of data [2]. From the last decades, data mining and knowledge discovery applications have important significance in decision making and it has become an essential component in various organizations and fields. The field of data mining has been increased day by day in the areas of human life with various integrations and advancements in the fields of Statistics, Databases, Machine Learning [3], Pattern Reorganization, Artificial Intelligence and Computation capabilities etc.

II. DATA MINING

According to the Gartner Group [4], "Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques."

"Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner" [5].

"Data mining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large data bases" [6].

III. WHY USE DATA MINING?

Data mining is to extract information from large amount of a data base. There are two main reasons to use data mining as a rapidly increase demands of data. These are:

- Too much data and too little information.
- There is a need to extract useful information from the data and to interpret the data.

IV. HISTORY OF DATA MINING

The term "Data mining" was introduced in the 1990s, but data mining is the evolution of a field with a long history [6].Data mining roots are traced back along three family lines: classical statistics, artificial intelligence, and machine learning:

- Statistics are the foundation of most technologies on which data mining is built, e.g. regression analysis, standard distribution, standard deviation, standard variance, discriminate analysis, cluster analysis, and confidence intervals. All of these are used to study data and data relationships.
- Artificial intelligence, or AI, which is built upon heuristics as opposed to statistics, attempts to apply human-thought-like processing to statistical problems. Certain AI concepts which were adopted by some high-end commercial products, such as query optimization modules for Relational Database Management Systems (RDBMS).
- Machine learning is the union of statistics and AI. It could be considered an evolution of AI, because it blends AI heuristics with advanced statistical analysis. Machine learning attempts to let computer programs learn about the data they study, such that programs make different decisions based on the qualities of the studied data, using statistics for fundamental concepts, and adding more advanced AI heuristics and algorithms to achieve its goals.

  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, AI, and machine learning. These techniques are then used together to study data and find previously-hidden trends or patterns within.

V. ISSUES OF DATA MINING

One of the key issues raised by data mining technologies is not a business or technological one, but social one. Some of the issues are address below:

A. Security and social issues

Today, Security [7] is an important issue with any data collection that is shared and/or is intended to be used for strategic decision-making. When data is collected for customer profiling, user behavior understanding, correlating personal data with other information, etc., large amounts of sensitive and private information about individuals or companies is gathered and stored. This becomes controversial given the confidential nature of some of this data and the potential illegal access to the information. Moreover, data mining could disclose new implicit knowledge about individuals or groups that could be against privacy policies, especially if there is potential dissemination of discovered information. Another issue that arises from this concern is the appropriate use of data mining. Due to the value of data, databases of all sorts of content are regularly sold, and because of the competitive advantage that can be attained from implicit knowledge discovered, some important information could be withheld, while other information could be widely distributed and used without control.

B. User interface issues

The knowledge discovered by data mining tools is useful as long as it is interesting, and above all understandable by the user. Good data visualization eases the interpretation of data mining results, as well as helps users better understand their needs. Many data exploratory analysis tasks are significantly facilitated by the ability to see data in an appropriate visual presentation. There are many visualization ideas and proposals for effective data graphical presentation. However, there is still much research to accomplish in order to obtain good visualization tools for large datasets that could be used to display and manipulate mined knowledge. The major issues related to user interfaces and visualization are “screen real-estate”, information rendering, and interaction. Interactivity with the data and data mining results is crucial since it provides means for the user to focus and refine the mining tasks, as well as to picture the discovered knowledge from different angles and at different conceptual levels.

C. Mining methodology issues

These issues pertain to the data mining approaches applied and their limitations. Topics such as versatility of the mining approaches, the diversity of data available, the dimensionality of the domain, the broad analysis needs (when known), the assessment of the knowledge discovered, the exploitation of background knowledge and metadata, the control and handling of noise in data, etc. are all examples that can dictate mining methodology choices. For instance, it is often desirable to have different data mining methods available since different approaches may perform differently depending upon the data at hand. Moreover, different approaches may suit and solve user’s needs differently

D. Performance issues

Many artificial intelligence and statistical methods exist for data analysis and interpretation. However, these methods were often not designed for the very large data sets data mining is dealing with today. Terabyte sizes are common. This raises the issues of scalability and efficiency of the data mining methods when processing considerably large data. Algorithms with exponential and even medium-order polynomial complexity cannot be of practical use for data mining. Linear algorithms are usually the norm. In same theme, sampling can be used for mining instead of the whole dataset. However, concerns such as completeness and choice of samples may arise. Other topics in the issue of performance are incremental updating, and parallel programming. There is no doubt that parallelism can help solve the size problem if the dataset can be subdivided and the results can be merged later. Incremental updating is important for merging results from parallel mining, or updating data

mining results when new data becomes available without having to re-analyze the complete dataset.

E. Data source issues

There are many issues related to the data sources, some are practical such as the diversity of data types, while others are philosophical like the data glut problem. We certainly have an excess of data since we already have more data than we can handle and we are still collecting data at an even higher rate. If the spread of database management systems has helped increase the gathering of information, the advent of data mining is certainly encouraging more data harvesting. The current practice is to collect as much data as possible now and process it, or try to process it, later. The concern is whether we are collecting the right data at the appropriate amount, whether we know what we want to do with it, and whether we distinguish between what data is important and what data is insignificant. Regarding the practical issues related to data sources, there is the subject of heterogeneous databases and the focus on diverse complex data types. We are storing different types of data in a variety of repositories. It is difficult to expect a data mining system to effectively and efficiently achieve good mining results on all kinds of data and sources. Different kinds of data and sources may require distinct algorithms and methodologies. Currently, there is a focus on relational databases and data warehouses, but other approaches need to be pioneered for other specific complex data types. A versatile data mining tool, for all sorts of data, may not be realistic. Moreover, the proliferation of heterogeneous data sources, at structural and semantic levels, poses important challenges not only to the database community but also to the data mining community

VI. TASKS OF DATA MINING

Fayyad et.al. (1996) defineS six main functions of data mining [8].These is the following:

A. *Classification :* Classification is finding models that analyze and classify a data item into several predefined classes.
B. *Sequencing :* Sequencing is similar to the association rule. The relationship exists over a period of time such as repeat visit to supermarket.
C. *Regression:* Regression is mapping a data item to a real-valued prediction variable.
D. *Clustering:* Clustering is identifying a finite set of categories or clusters to describe the data.
E. *Dependency Modeling:* Dependency Modeling (Association Rule Learning) is finding a model which describes significant dependencies
F. between variables.
G. *Deviation Detection:* Deviation Detection (Anomaly Detection) is discovering the most significant changes in the data.
H. *Summarization:* Summarization is finding a compact description for a subset of data.

VII. DATA MINING APPLICATIONS

Data mining is a data analysis approach that has been quickly adapted and used in a large number of domains that were already using statistics. The applications areas of data mining are :

Medical / Pharmacy

- Computer Assisted Diagnosis (expert systems learning).
- Characterization/prediction of patient's response to product dosage.
- Identification of successful medical therapies (successful prescription patterns).
- Study of relations between dosage and potentially related adverse events.

B. Insurance and Health Care

- Discovery of medical procedures that are claimed together through claims analysis
- Identification of customers that are potential buyers for new policies.
- Detection of behavior patterns capable of identifying risky customers.

Detection of *Banking / Finance*

- Detection of fraudulent credit card usage patterns.
- Risk management related to attribution of loans using scorecards.
- Find hidden correlations between different financial indicators.
- Identification of stocks trading rules from historical market data *Retail / Marketing* Discovery of buying behavior patterns
- Detection of associations among customer characteristics.
- Prediction of the probability that clients answer to mailing.

VIII. CONCLUSIONS

Data mining is to discover or extract knowledge or data from large amount of database. In this paper, introduce briefly reviewed the concept of data mining, issues of data mining and areas of data mining where used today. It would be helpful to researchers to focus on the various issues and challenges of data mining. 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. From the last decades, data mining and knowledge discovery applications have important significance in decision making and it has become an essential component in various organizations and fields.

IX. FUTURE TRENDS

The complexity of data mining must be hidden from end-users before it will take the true center stage in an organization. Business use cases can be designed, with tight constrains, around data mining algorithms. Due to the enormous success of various application areas of data mining, the field of data mining has been establishing itself as the major discipline of computer science and has shown interest potential for the future developments.

REFERENCES

[1] Gorunescu, F, Data *Mining: Concepts, Models, and Techniques*, Springer, 2011.
[2] Han, J., and Kamber, M. , *Data mining: Concepts and techniques*, Morgan-Kaufman Series of Data Management Systems San Diego:Academic Press, 2001.
[3] Neelamadhab Padhy, Dr. Pragnyaban Mishra and Rasmita Panigrahi, "The Survey of Data Mining Applications and Feature Scope, *International Journal of Computer Science, Engineering and Information Technology* (IJCSEIT)", vol.2, no.3, June
[4] Heikki, Mannila, *Data mining: machine learning, statistics and databases*, IEEE, 1996.
[5**]** Fayadd, U., Piatesky -Shapiro, G., and Smyth, P, From Data Mining To Knowledge Discovery in Databases", The MIT Press, ISBN 0–26256097–6, Fayap, 1996.
[6**]** Piatetsky-Shapiro, Gregory, *The Data-Mining Industry Coming of Age*,"IEEE Intelligent Systems, 2000.
[7] Jing He, *Advances in Data Mining: History and Future*, Third international Symposium on Information Technology Application, 978-0-7695- 3859-4, IEEE, 2009.
[8] Berry, M.and Linoff, G., *Master Data Mining: The Art and Science of Customer Relationship Management*, Wiley publisher, 2000.