Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 4, April 2014, pg.1401 – 1406

REVIEW ARTICLE

A Review of Data Mining Techniques

Sukhdev Singh Ghuman

Dept. of Comp. Sci., SBDSM Khalsa College, Domeli (Kapurthala), Punjab, India

ghumanggg@gmail.com

Abstract- Information technology has revolutionized the whole world with cheaper and fast communication through different modes. All these devices generate lots of data which need to be processed to extract useful patterns of data or information. The database technologist are seeking means to store, manipulate and retrieve data while data mining area is striving hard to find new and efficient techniques for information extraction form the vast amount of data. Data Mining is also referred by the names like Knowledge Discovery in Database (KDD) or Predictive Analytics or Data Science. The various techniques used for extraction are genetic algorithms, decision trees, artificial neural networks, induction and visualization. Data mining is generally an iterative and interactive discovery process. The goal of this process is to mine patterns, associations, changes, anomalies, and statistically significant structures from large amount of data.

Keywords: Data Mining, Patterns, Knowledge Discovery, Database, Techniques

I. INTRODUCTION

Data mining is the process of discovering patterns in the large data sets. The purpose of the data mining is to find information from the large data sets and convert it into usable structures so that this information can used for further processing without any difficulty. It is handled by databases and managed by database management aspects. This is a commonly used word for any kind of large scale data processing. The term data mining was discovered around 1990 in computer science. It is also referred by several other terms like Knowledge Discovery in Databases (KDD) or Predictive Analytics or Data Science [1]. Data mining is generally an iterative and interactive discovery process. The goal of this process is to mine patterns, associations, changes, anomalies, and statistically significant structures from large amount of data [3]. The mined results should be valid, novel, useful, and understandable. This paper presents a brief introduction about data mining in section 1. The second section illustrates the process of data mining while the third

section reviews different data mining techniques. The fourth section is committed to Knowledge Discovery in Databases (KDD) and fifth section discusses some issues relating to data mining. The last section presents the conclusion.

II. PROCESS OF DATA MINING

The process of data mining is sequential which require many steps to be followed which are as shown below in the form of a diagram [3].

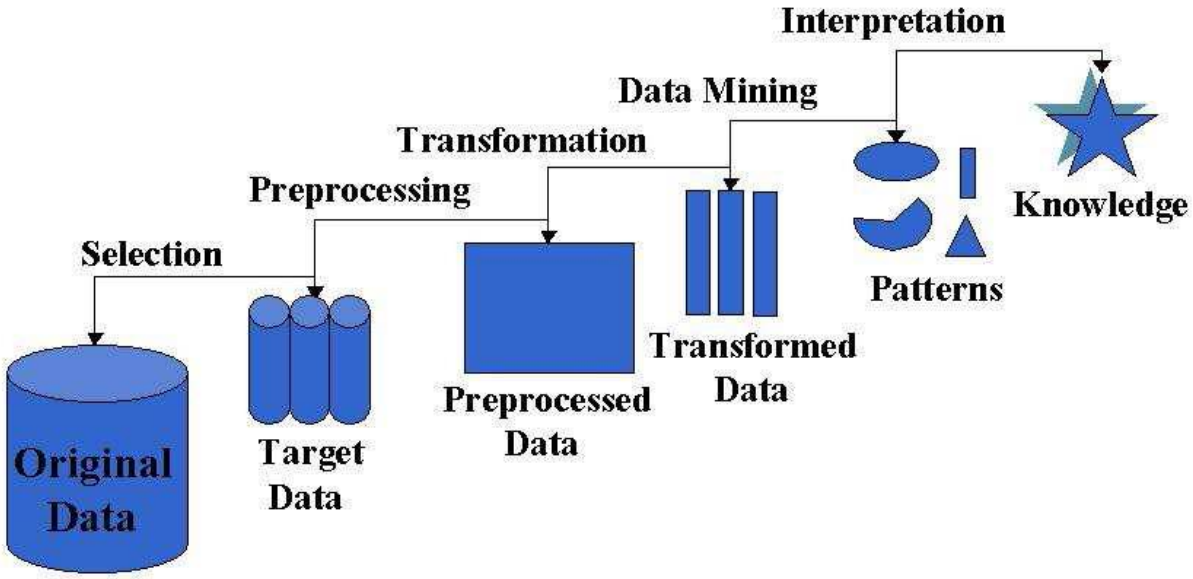

Figure 1: Data mining process [3]

1. Extract, transform, and load transaction data onto the data warehouse system.
2. Store and manage the data in a multidimensional database system.
3. Provide data access to business analysts and information technology professionals.
4. Analyze the data by application software.
5. Present the data in a useful format, such as a graph or table.

III. TECHNIQUES OF DATA MINING

Data mining is complex process and it requires not only fast processing devices but good and efficient techniques of data processing. The important techniques of data mining are as listed below:-.

- **Artificial neural networks**: AI techniques are widely used in Data Mining. Techniques such as pattern recognition, machine learning, and neural networks are very useful. Many

other techniques in AI such as knowledge acquisition, knowledge representation, and search, are relevant to the various process steps in data mining [4]. It is a non-linear predictive model. It learns through training and resembles biological neural networks in structure.

- **Genetic algorithms**: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution. It is a relatively new software paradigm inspired by Darwin's theory of evolution. A population of rules, each representing a possible solution to a problem, is initially created at random. Then pairs of rules are combined to produce off spring for the next generation. A mutation process is used to randomly modify the genetic structures of some members of each new generation. The system runs for dozens or hundreds of generations. The process is terminated when an acceptable or optimum solution is found, or after some fixed time limit. Genetic algorithms are appropriate problems that require optimization with respect to some computable criterion. This paradigm can be applied to Data Mining problems. Large and complex problems require a fast computer in order to obtain appropriate solutions in a reasonable amount of time. Mining large data sets by genetic algorithms has become practical only recently due to the availability of affordable high-speed [4].

- **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). These are decision tree techniques used for classification of a dataset. They provide a set of rules that can be applied to a new dataset to predict which records will have a given outcome [4].

- **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**: This technique is used for the extraction of useful if-then rules from data based on statistical significance.

- **Data visualization**: It is concerned with visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships. A picture is worth thousands of numbers. Visual data mining techniques have proven the value in exploratory data analysis, and they also have a good potential for mining large database. This approach requires the integration of human in the data mining process.

IV. KNOWLEDGE DISCOVERY IN DATABASES

Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps [5]:

- **Data cleaning**: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.
- **Data integration**: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.
- **Data selection**: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.
- **Data transformation**: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
- **Data mining**: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.
- **Pattern evaluation**: in this step, strictly interesting patterns representing knowledge are identified based on given measures.
- **Knowledge representation**: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

V. ISSUES IN DATA MINING

Data mining algorithms embody techniques that have sometimes existed for many years, but have only lately been applied as reliable and scalable tools that time and again outperform older classical statistical methods. While data mining is still in its infancy, it is becoming a trend and ubiquitous. Before data mining develops into a conventional, mature and trusted discipline, many still pending issues have to be addressed. Some of these issues are addressed below. Note that these issues are not exclusive and are not ordered in any way [2].

- **Security and Social Issue:** Security is an important issue with any data collection that is intended to be shared. It is the issue of individual privacy. Data mining makes it possible to analyze routine business transactions and glean a significant amount of information about individuals buying habits and preferences.
- **Data integrity:** Data analysis can only be as good as the data that is being analyzed. A key implementation challenge is integrating conflicting or redundant data from different sources. For example, a bank may maintain credit cards accounts on several different databases. The addresses (or even the names) of a single cardholder may be different in each. Software must translate data from one system to another and select the address most recently entered.
- **Mining Methodology:** An important technical issue is whether it is better to set up a relational database structure or a multidimensional one. In a relational structure, data is stored in tables, permitting ad hoc queries. In a multidimensional structure, on the other hand, sets of cubes are arranged in arrays, with subsets created according to category. While multidimensional structures facilitate multidimensional data mining, relational structures thus far have performed better in client/server environments. And, with the explosion of the Internet, the world is becoming one big client/server environment.
- **Cost:** Finally, there is the issue of cost. While system hardware costs have dropped dramatically within the past five years, data mining and data warehousing tend to be self-reinforcing. The more powerful the data mining queries, the greater the utility of the information being gleaned from the data, and the greater the pressure to increase the amount of data being collected and maintained, which increases the pressure for faster, more powerful data mining queries. This increases pressure for larger, faster systems, which are more expensive [5].
- **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.

VI. CONCLUSION

Data mining is concerned with extracting useful rules or interesting patterns from the bulk amount of data collected through various sources. There are many data mining techniques which can be used to perform the job efficiently. It is to be noted that a single technique cannot be used for all types of data because depending on the type of data, appropriate technique is available for extraction of information. Sometimes hybrid techniques are more useful instead of a single technique.

REFERENCES

[1] http:/www.wikipadeia.com

[2] http://www.anderson.ucla.edu

[3] Mohammed J. Zaki, "DATA MINING TECHNIQUES", August 2003

[4] Sang Jun Lee, Keng Siau, "A Review of Data Mining Techniques" Industrial Management & Data Systems 101/1 (2001) 41-46

[5] Dr. Rajni Jain, "Introduction to Data Mining Techniques"

