ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 10, March 2018

Data Mining Processes and Techniques

Virpal Kaur

Assistant Professor in Computer Science and Applications Swami Ganga Giri Janta Girls College, Raikot (Ludhiana), Punjab,India Email ID-virpal.kaler@rediffmail.com

Abstract: Data and information has a significant role on human activities. Data mining process analyses the large volume of data from various perspectives and summarizing it into valuable information. This paper conducts a formal review of data mining processes. It also provides a survey of various data mining techniques. These techniques include association, classification, clustering, prediction and sequential patterns. This paper discusses the topic based on past research papers.

Keywords: Data Mining, Data Mining Processes, Data Mining Techniques.

I. INTRODUCTION

Data mining is defined as extracting or mining information from large quantity of data which is stored in multiple heterogeneous data base such as data warehouse, external sources and using it to make crucial business decisions. In other words, we can say that data mining is mining the knowledge from data that improves business efficiency. The term data mining is also treated as Knowledge Mining from Data, Knowledge Discovery in the Database (KDD) or Knowledge Mining. It has become possible for organizations to conglomerate large amount of data at lower cost due to data collection and storage technology. The overall goal of data mining is exploiting this stored data in order to extract useful and actionable information and also to discover meaningful patterns and rules and transform it into an understandable structure for further use. Intelligent methods are applied. Problems are solved by analyzing data already present in databases. It is a type of sorting technique which is actually used to extract hidden patterns from large database and those patterns which are previously not explored.

II. ELEMENTS OF DATA MINING

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.

It is an iterative process, which consist of following steps shown in Fig.1.

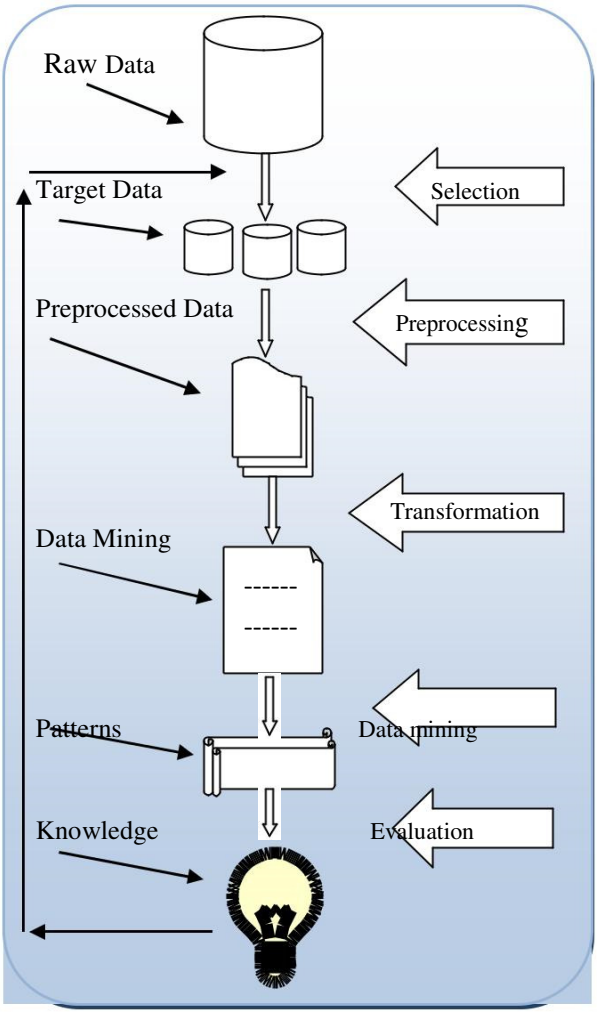

Fig 1 knowledge discovery process

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 10, March 2018

- Data comes from variety of sources is integrated into a single data store called target data.
- Then data is pre-processed and transformed into standard format.
- The data mining algorithms process the data to the output in form of patterns or rules.
- Then those patterns and rules are interpreted to new or useful knowledge or information.
- The ultimate goal of this process is to find the hidden patterns and interpret them to useful knowledge and information.

III. DATA MINING PROCESSES

Data mining consists of various numbers of processes because the whole process cannot be accomplished in a single step and it must be reliable. Therefore, to reduce its complexities the entire process is divided into number of following different steps as shown in Fig.2.

Data Cleaning: Data cleaning is a process which ensures the completion and consistency of data. The data obtained from real world is normally incomplete, noisy and in some situations data contains errors or outliers. At times the availability of data might be lacking attribute values, data of interest etc. Hence, the data mining results would be neither reliable nor accurate. To solve this problem data cleaning process is followed.

Data Integration: Data integration is a processing technique that merges the data from multiple heterogeneous data sources into a coherent data store and provides an unfiled view of data to its users. Data integration is a really complex and tricky task because data gathered from different sources like data base, text files does not match normally. Despite the inconsistency of data another issue is redundancy. Data integration attempts to reduce the redundancy of data to the maximum possible level. Fig. 2

Data Selection: The large volume of historical data is collected for analysis purpose. But the available data have much more data than the actual requirement. So, in this step irrelevant information is abstracted to gain only relevant data.

Data Transformation: Data transformation is the process where transformation and consolidation of data is done. It includes the conversion of data from one structure to another structure that is most appropriate for data mining. It can be divided into the following steps:

- Data discovery
- Data mapping
- Code generation
- Code execution
- Data review

Data Mining: The next step of data mining is the core process when transformed data is mined by using various complex techniques to extract patterns of interest and the integrated data to enable easy identification of any valuable information.

Pattern Evaluation: In this step, an evaluation is made to find out the most interesting patterns based on their potential usefulness, understandability, validation and confirmation on new data with some degree of certainty.

Knowledge Representation: This step consist of presenting the output of data mining to the users by applying different representations and visualization techniques like graphs, histogram, tables etc.

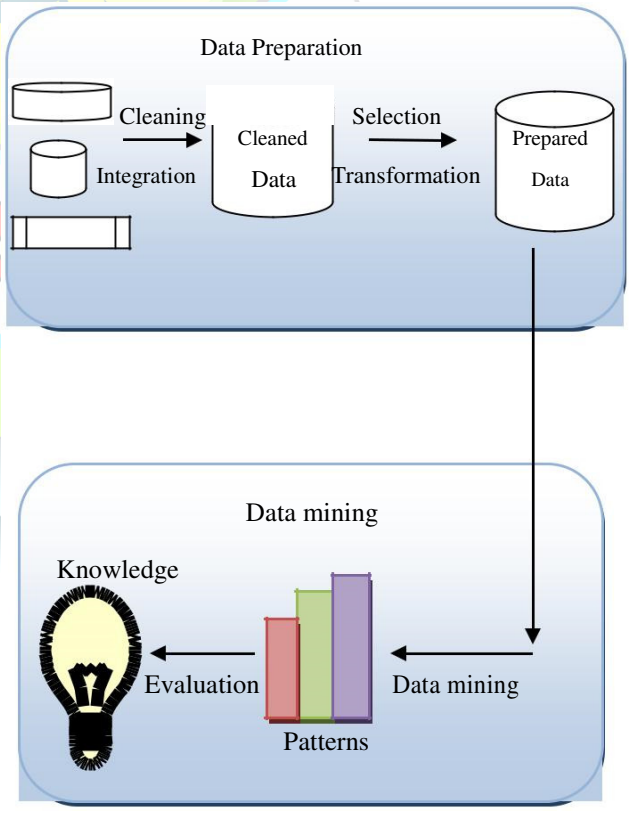

Fig. 2 Data mining processes

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 10, March 2018

IV. DATA MINING TECHNIQUES

Various algorithms and techniques, which are used for knowledge discovery from databases, are association, classification, clustering, prediction and sequential patterns etc.

Association: Association is one of the most useful techniques in data mining. This technique discovers a pattern to find out the relationship of two items involves in same transaction. In other words, association finds the probability of the co-occurrence of items in a collection. The relationships between co-occurring items are expressed as association rules. For example, association technique is used to study the customer behavior.

For example, when tracking customer' buying behavior retailers might identify that a female customer always buy jewelry items when they buy dresses, and therefore suggest that the next time they buy dresses they might also want to buy jewelry.

Association rule offers two major information:

- Support – How often is the rule applied?
- Confidence – How often the rule is correct?

This technique follows a two step process

- Find all the frequently occurring data sets
- Create strong association rules from the frequent data sets

.

There are various types of association rule:

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

Clustering: Clustering is a process which finds out similarities among data based on their characteristics. It is a data reduction technique that involves grouping of data where one group is heterogeneous to other groups while having homogeneity within it. It is based on unsupervised learning. Clustering helps to recognize the differences and similarities between the data. Clustering is also defined as segmentation that helps the users to figure out what is happening within the database.

For example, a marketing company can group its customers based on their income, age, gender, education level, occupation and family size. There are various types of clustering methods:

- Partitioning Methods
- Density Based Methods
- Grid Based Methods
- Hierarchical Agglomerative Methods
- Model Based Methods

Classification: Classification technique comes under supervised learning where desired output for a given input is known and it is based on machine learning. Items in a set of data are classified into one of predefined set of classed based on the training set and values. This technique is useful for predicting group membership for data instances. The future of data can be predicted. Important information can be derived about data and metadata (data about data). Different techniques are used by different classification algorithms for fining the relationship. Classification has many applications in business modeling, credit analysis, marketing and biomedical. The goal of classification is to accurately predict the target class for each case in the data. There are two main processes involved in this technique:

- Learning: A Classification algorithm is used to analyze the data.
- Classification: The precision of the classification rules is measured by using the data.

For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks. There are various types of classification:

- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations

Prediction: This technique is used to find the relationship between dependent and independent variables. Prediction model is based on continuous or ordered values. Prediction in data mining is to identify data points purely on the description of another related data values. It is used to identify the data points on the description of another related data values and it is not necessarily related to future events but the used variables are unknown. It is used to find out the relationship between a thing you know and a thing you need to predict for future reference.

For example, prediction model is used by marketing manager to predict the maximum amount spent by a particular customer so that the upcoming sale amount can be planned.

Sequential Patterns: Sequential data is omnipresent. This technique is used to analyze this data and identify patterns and efficient systems are

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 10, March 2018

implemented by these patterns. Sequential pattern is used to find subsequences that appear often are common to several sequences and those subsequences are called the frequent sequential patterns.

All of these techniques can help to analyze different data from different perspectives. Data mining techniques are very useful for companies to analyze their big data in effective way. The decision of selecting the appropriate technique is based on the nature of problem, availability of data or the desired results. No single technique can be used to solve the problems.

V. CONCLUSION

This paper provides a general idea of data mining, data mining processes and its various techniques. There are also various applications of data mining in real life including biomedical, DNA data analysis, telecommunication and retail industry, education sector, marketing, manufacturing, fraud detection, and financial data analysis. Future research will involve the development of new techniques for incorporating uncertainty management in data mining.

REFERENCES

1. http://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining
2. http://www.wideskills.com/data-mining-tutorial/data-mining-processes
3. http://www.zentut.com/data-mining/data-mining-processes/
4. www.tutorialride.com/data-mining/knowledge-representation-in-data-minibg.htm
5. www.ibm.com/developerworks/library/ba-data-mining-techniques/
6. www.lifewire.com/classification-1019653
7. www.quora.com/What-is-prediction-in-data-mining
8. Data-mining.philippe-fournier-viger.com/introduction-sequential-pattern-mining/
9. www.wisdomjobs.com/e-university/data-mining-tutorial-199/data-mining-techniques-1872.html
10. www.educba.com/7-data-mining-technique-for-best-results/
11. Kumar, Puneet, Bhardwaj, Sushil (2017). Advanced database Programming and MySQL. India. Kalyani Publisher

BIOGRAPHY

Author Name: Ms. Virpal Kaur
Designation: Assistant Professor
Department: Post Graduate Department of Computer Science and Applications
Qualification: B.C.A., M.C.A.