Research Paper Information Technology

Review of Data Mining: Techniques, Applications and Issues

*Keyur J Patel

*Master of Technology, Dept. Information Technology, U V Patel College of Engineering, Kherva, Mehsana, India

ABSTRACT

Data or Knowledge has an important role on human activities. Data mining is the knowledge discovery process by analyzing the large volumes of data from various a particular way of seeing information and a brief statement of the main points it into useful information. In this paper we have focused a variety of techniques, approaches and different areas of the research which are helpful in data mining and its applications. The huge amount of data is available in the form of tera- to peta-bytes. To analyze, manage and make a decision of such type of huge amount of data we need techniques. This paper gives the knowledge of the data mining and some of its applications and also focuses on scope of the data mining.

Keywords : Data Mining Techniques, Applications and Issues

INTRODUCTION

To generate information it requires massive collection of data. The data can be simple numerical figures and text documents, to more complex information such as spatial data, multimedia data, and hypertext documents. As the data are available in the different formats so that the proper action to be taken. Not only to analyze these data but also take a good decision and maintain the data. This technique is actually we called as a data mining.

Data mining is the extraction of hidden predictive information from large databases [1] [3]. The core functionalities of data mining are applying various methods and algorithms in order to discover and extract patterns of stored data. The complete data mining process is a combination of many sub processes. Some important are data extraction, data cleaning, feature selection, algorithm design, tuning and analysis of the output when the algorithm is applied to the data.

The field of data mining have been prospered and posed into new areas of human life with various integrations and advancements in the fields of Statistics, Database, Machine Learning, Pattern Reorganization, Artificial Intelligence and Computation capabilities etc. Here is an overview of the data mining techniques and some of its applications and also focus on scope of the data mining.

DATA MINING LIFE CYCLE

The life cycle of a data mining project consists of six phases [2][4]. The sequence of the phases is not rigid. The main phases are:

A. Business Understanding:
This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

B. Data Understanding:
It starts with an initial data collection, to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

C. Data Preparation:
It covers all activities to construct the final dataset from the initial raw data.

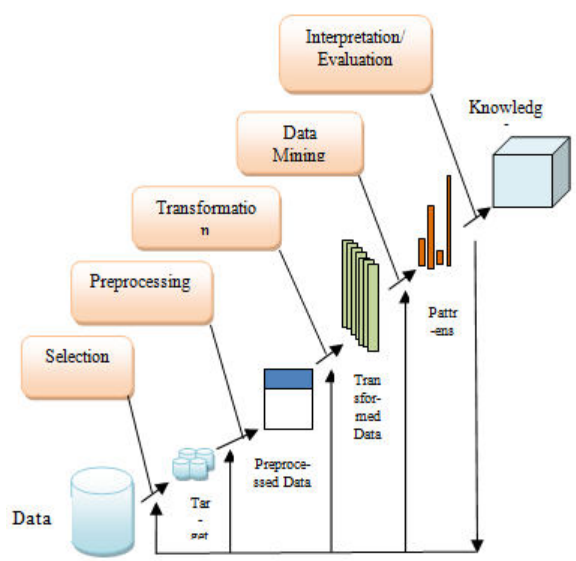

Figure 1: Data Mining Life Cycle

D. Modeling:
In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.

E. Evaluation:
In this stage the model is thoroughly evaluated and reviewed. At the end of this phase, a decision on the use of the data mining results should be reached.

F. Deployment:
The purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. The deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

DATA MINING TECHNIQUES

There are several major data mining techniques have been developed and used in data mining [8]:

A. Association

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction. For example, the association technique is used in market basket analysis to identify what products that customers frequently purchase together. Based on this data businesses can have corresponding marketing campaign to sell more products to make more profit.

B. Classification

Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. For example, we can apply classification in application that "given all past records of employees who left the company, predict which current employees are probably to leave in the future." In this case, we divide the employee's records into two groups that are "leave" and "stay".

C. Clustering

Clustering is a data mining technique that makes meaningful or useful cluster of objects that have similar characteristic using automatic technique. Different from classification, clustering technique also defines the classes and put objects in them, while in classification objects are assigned into predefined classes. For example in a library, books have a wide range of topics available. The challenge is how to keep those books in a way that readers can take several books in a specific topic without hassle. By using clustering technique, we can keep books that have some kind of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in a topic, he or she would only go to that shelf instead of looking the whole in the whole library.

D. Prediction

The prediction as it name implied is one of a data mining techniques that discovers relationship between independent variables and relationship between dependent and independent variables. For instance, prediction analysis technique can be used in sale to predict profit for the future if we consider sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction.

E. Sequential Patterns

Sequential patterns analysis in one of data mining technique that seeks to discover similar patterns in data transaction over a business period. The uncover patterns are used for further business analysis to recognize relationships among data.

DATA MINING APPLICATIONS

The data mining applications can be generic or domain specific. The generic application is required to be an intelligent system that by its own can takes certain decisions like: selection of data, selection of data mining method, presentation and interpretation of the result. Some generic data mining applications cannot take its own these decisions but guide users for selection of data, selection of data mining method and for the interpretation of the results [5][6].

A. In **M**edical **S**cience there is large scope for application of data mining. Diagnosis of diesis, health care, patient profiling and history generation etc. are the few examples.
B. **P**erforming **B**asket Analysis- Also known as affinity analysis, basket analysis reveals which items customers tend to purchase together. This knowledge can improve stocking, store layout strategies, and promotions.
C. **S**ales **F**orecasting- Examining time based patterns helps retailers make stocking decisions. If a customer purchases an item today, when are they likely to purchase a complementary item?
D. The classification method of data mining is used to classify the **N**etwork **T**raffic normal traffic or abnormal traffic.
E. In the sports world the vast amounts of statistics are collected for each player, team, game, and season. Data mining can be used by **S**ports **O**rganizations in the form of statistical analysis, pattern discovery, as well as outcome prediction.
F. Data mining is used for **I**dentifying **P**atterns that characterize successful groups from less successful ones. The data mining algorithms are used that can properly account for the temporal nature of the data and the character of group interaction.
G. In **W**eb-based **E**ducation the data mining methods are used to improve courseware. The relationships are discovered among the usage data picked up during students' sessions. This knowledge is very useful for the teacher or the author of the course, who could decide what modifications will be the most appropriate to improve the effectiveness of the course.
H. The data mining algorithms effectively used for **P**rediction of the **P**ersonal **B**ankruptcy. The data mining method least squares regression; neural nets and decision trees are proved to be the suitable for prediction of bankruptcy.

ISSUES IN DATA MINING

Before data mining develops into a conventional, mature and trusted discipline, many still pending issues have to be addressed which are [7]:

A. Security and Social Issues:

When data is collected from user 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.

B. User Interface Issues:

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:

Most algorithms assume the data to be noise-free. Most datasets contain exceptions, invalid or incomplete information, etc. As a consequence, data pre-processing becomes time-consuming and frustrating.

D. Performance Issues:

When processing large data sets raises the issues of scalability and efficiency of the data mining methods and incremental updating, and parallel programming.

E. Data Source Issues:

It includes diversity of data types. We 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.

CONCLUSION

Data mining has importance regarding finding the patterns, forecasting, discovery of knowledge etc. in different business domains. Data mining applications use the variety of data types, range from text to images and stores in variety of databases and data structures. The different methods of data mining are used to extract the patterns and thus the knowledge from this variety databases. Selection of data and methods for data mining is an important task in this process and needs the knowledge of the domain. For data mining, variety of data that should be collected in the specific problem domain, selection of specific data for data mining, cleaning and transformation of data, extracting patterns for knowledge generation and fi-

nally interpretation of the patterns and knowledge generation. Data mining is used in medical science, detect malicious executables, sports organizations, identifying patterns, sales forecasting, performing basket analysis etc. Before data mining develops into a conventional, mature and trusted discipline, many still pending issues like security and social issues, user interface issues, performance issues etc.

REFERENCES

[1] Dunham, M. H., Sridhar S., "Data Mining: Introductory and Advanced Topics", Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006 | [2] Larose, D. T., "Discovering Knowledge in Data: An Introduction to Data Mining", ISBN 0-471-66657-2, ohn Wiley & Sons, Inc, 2005. | [3] Fayyad, U., Piatetsky-Shapiro, G., and Smyth P., "From Data Mining to Knowledge Discovery in Databases," AI Magazine, American Association for Artificial Intelligence, 1996. | [4] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R... "CRISP-DM 1.0 : Step-by-step data mining guide, NCR Systems Engineering Copenhagen (USA and Denmark), DaimlerChrysler AG (Germany), SPSS Inc. (USA) and OHRA Verzekeringenen Bank Group B.V (The Netherlands), 2000". | [5]Mr. S. P. Deshpande1 and Dr. V. M. Thakare2, "Data Mining System And Applications: A Review", IJDPS, Vol.1, No.1, September 2010, pp.36-41. | [6]Neelamadhab Padhy, Dr. Pragnyaban Mishra , and Rasmita Panigrahi, "The Survey of Data Mining Applications And Feature Scope", IJCSEIT, Vol.2, No.3, June2012,pp.48-53. | [7]Data Mining Techniques available at | http://www.zentut.com/data-mining/data-mining-techniques/%E2%80%9D