ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 6, June 2013

A View on Data Mining

Manpreet Kaur Mand[1], Gunjan[2], Diana Nagpal[3]

Assistant Professor, CSE, GNDEC, Ludhiana, India [1, 2, 3]

Abstract: Data mining, or knowledge discovery, is the computer-assisted process of analyzing voluminous sets of data and then discovering the meaning of the data. Data Warehouse is a subject-oriented, time variant, integrated, non volatile collection of data to assist management in the decision making process Data mining tools predict behaviours and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were very time consuming to resolve. They evaluate databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. This paper also shows the architecture of data mining system. Generally, Data mining tasks can be classified into two main categories namely descriptive and predictive. Descriptive Mining tasks characterize the general properties of data in the database while Predictive Mining tasks inference on the current data to make predictions. Class Description and Associations are the activities under descriptive mining and Regression and cluster analysis are the tasks under predictive mining. Data mining systems must be able to discover patterns at various levels of abstraction. Data mining systems also facilitates users to provide certain hints to guide the search for required patterns.

Keywords: Data Mining, Data Warehousing, Knowledge Discovery in databases, cluster analysis, class description.

I. INTRODUCTION

Due to wide availability of huge amounts of data and need for converting such data into useful information and knowledge, data mining concepts and techniques are used. Data mining refers to the process of extracting or uncovering the hidden patterns in the large data sets. Data mining is an integral step of KDD (Knowledge Discovery in databases). KDD as a process is illustrated in the Fig. 1 and consists of iterative sequence of following steps [1]:

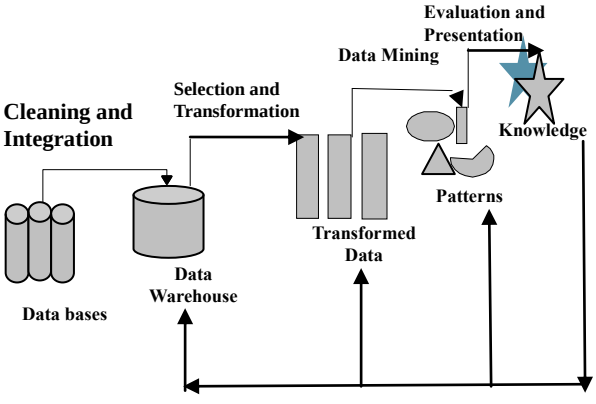

Fig.1. Data mining as a step in the process of KDD

1. Data Cleaning: To remove inconsistent, irrelevant, missing values and noisy data from data base.
2. Data Integration: To combine data from multiple databases.
3. Data Selection: Data relevant to the current problem analysis is retrieved from the database.
4. Data Transformation: Data is converted into a form appropriate for data mining.
5. Data Mining: A process where algorithms are applied to extract the unseen data patterns.
6. Pattern Evaluation: To identify truly interesting patterns representing knowledge.
7. Knowledge Presentation: Mined Knowledge is presented to the user by using certain knowledge presentation techniques.

II. DATA WAREHOUSE

Data Warehouse is a subject-oriented, time variant, integrated, non volatile collection of data to assist management in the decision making process. A data warehouse must include all the data available to an organisation regardless of the location of data.

Data warehouse is the huge repository collected from multiple locations stored under a unified schema and that normally reside at a single site. Data warehouses are constructed via a process of data cleaning, data integration, data transformation and periodic data refreshing [12]. Data Warehouses are used to store both detailed data as well as summarized data. Detailed data is used in the process of pattern analysis and summarized data are used to hold the

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 6, June 2013

results of historical analysis [3]. Effective and efficient organisation of Data Warehouse is the essential part of efficient data mining.

According to Inmon, a Data warehouse the following characteristics [4]:

1. Subject Oriented: The data are stored according to the specific area of business or specific aspect of the company interest.

2. Integrated: Data is integrated from variety of sources after identifying and correcting inconsistencies.

3. Provides support for Management decisions.

4. Variant in Time: The data consists of a series if sophisticated snapshots obtained at a particular moment of time.

III. ARCHITECTURE OF DATA MINING SYSTEM

Architecture of typical data mining system has following major components shown as Fig. 2[1]:

1. Database, Data Warehouse, World Wide Web or other information repository.
2. Database or Data Warehouse Server.
3. Knowledge Base.
4. Data Mining Engine.
5. Pattern Evaluation Module.
6. User Interface.

Database, Data Warehouse, World Wide Web or other information repository: It is a set of databases, data warehouses and/or other kinds of information repositories on which data cleaning and data integration techniques are employed.

Database or Data Warehouse Server: It is responsible for fetching the relevant data based on the user's requirement.
Knowledge Base: It is the domain Knowledge that is used to assist the search and identify the hidden patterns.

Data Mining Engine: It is the inherent part of the data mining system and constitutes a set of modules for activities like characterization, association and correlation analysis, classification, prediction and cluster analysis, outlier analysis and evolution analysis.

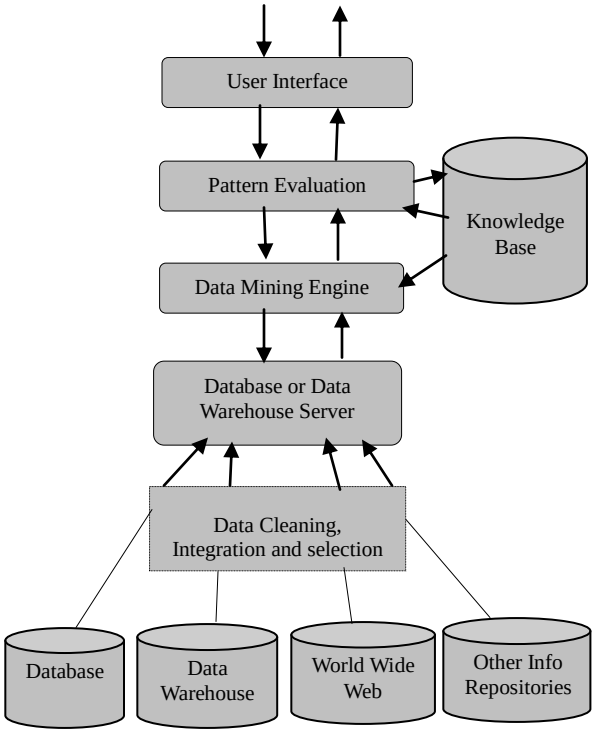

Fig.2. Architecture of Data mining System

Pattern Evaluation Module: This component typically employs interestingness measures and interacts with the Data Mining module to explore the search for getting interesting hidden patterns. Alternatively the pattern evaluation module works in integration with the data mining module.

User Interface: This module provides an interface between the user and data mining system, allows user to interact with the system by specifying a data mining query, provide information regarding the search, and performing exploratory data mining based on intermediate data mining results.

IV. DATA MINING TECHNIQUES

Generally, Data mining tasks can be classified into two main categories namely descriptive and predictive. Descriptive Mining tasks characterize the general properties of data in the database while Predictive Mining tasks inference on the current data to make predictions. Descriptive tasks summarize the data in some manner. Examples of such tasks include automatically segmenting customers based on their similarities and differences [5] and finding associations between products in market basket data [6]. On the other hand, Predictive tasks allow one to predict the value of a variable based on other existing information. Examples of predictive tasks include predicting when a customer will leave a company [7], predicting whether a transaction is fraudulent or not [8], and identifying the best

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 6, June 2013

customers to receive direct marketing offers [9]. Data mining systems must be able to discover patterns at various levels of abstraction. Data mining systems also facilitates users to provide certain hints to guide the search for required patterns.

A. Concept/Class Description: Characterization and Discrimination

Data can be associated with classes or concepts. These descriptions can be derived via data characterization, data discrimination or both data characterization and data discrimination. Data Characterization is the summarization of the data of a class under study (called Target class). Data corresponding to user specified class is generally collected as database query. Data discrimination is the comparison of general features of target class data objects with the general features of the objects from one or more comparative classes (called contrasting classes). Discrimination descriptions are expressed in the form of rules called discriminant rules.

B. Mining frequent patterns, associations and correlations

Frequently occurred patterns in data are called frequent patterns. Mining those frequent patterns results into extracting interesting associations and correlations which are hidden in data so far. Association Analysis is a descriptive Data Mining Technique to uncover pattern or associations between data [10]. Rules or implications are used to represent associations. One possible association rule from Supermarket data is Buy (computer) → Buy (software) indicates that is a customer buys a computer then he or she will also buys software as well. Correlation is a basic technique to find relationship between variables. It is a measure of linear relationship between two variables [11]. Pearson's correlation coefficient is given by:

$$\text{corr}(x,y) = \frac{\sum_{i=1}^{n} (x_i - x')(y_i - y')}{\left[\sum_{i=1}^{n} (x_i - x')^2 \sum_{i=1}^{n} (y_i - y')^2 \right]}$$

C. Classification and Regression

These are the predictive tasks that involves to the building of model to predict a target or dependent variable from a set of explanatory or independent variables [10]. For classification tasks the target value normally has a small number of discrete values i.e. high and low whereas for regression tasks the target variable is continuous.

D. Cluster Analysis

In cluster analysis task, the objective is to group similar (homogeneous) objects in the same cluster and dissimilar (heterogeneous) objects in the different clusters [10]. With customer clustering you are able to find the buying habits of customers and then take the advantage of these similarities to target offers to only those subgroups.

V. CONCLUSION

Data mining is the depth analysis of raw data to find some kind of trend or usable information that will aid in decision making for a business or government. Data mining is a powerful technology with great potential to assisting companies in their decision making process. It is a crucial part of the process of knowledge discovery in databases. Two techniques of data mining have been discussed in this paper i.e. Descriptive Mining and Predictive Mining. This paper also explains the architecture of typical data mining system.

ACKNOWLEDGMENT

Authors are highly grateful to Mr. Amanpreet Singh Brar, Associate Professor and Head, CSE Department, GNDEC, Ludhiana for his precious guidance from time to time.

REFERENCES

J.Han, M.Kambe and J.Pei, *Data Mining: Concepts and Techniques*, 2nd ed., Diane Cerra, San Fransisco, CA, 2006.

S.Hober (1997),"Understanding The Data Warehouse Process," [online].Available:http://www.ukpua.org/archives/articles/warehouse.html

[3] S P. Imberman , "Effective Use Of The Kdd Process And Data Mining For Computer Performance Professionals, " In proc. Int. CMG 2001,pp. 611-620.

[4] W.H.Inmon , *Building the Data Warehouse* , 2nd ed.,Katherine Schowalter,Canada,John Wiley and Sons,Inc.,1996.

[5] Y.Chen,, G.Zhang, D.Hu, and S.Wang, Customer Segmentation in customer relationship management based on data mining. In *Knowledge Enterprise: Intelligent Strategies in Product Design, Manufacturing, and Management*, pp.288-293. Boston: Springer.2006.

[6] R Agrawal., and R.Srikant., Fast algorithms for mining association rules. In *Proc. International Conference on Very Large Databases*, 1994, pp.487- 499.

[7] C.Wei, and I.Chiu, Turning telecommunications call details to churn prediction: A data mining approach. *Expert Systems with Applications*, vol. 23 (2), pp.103-112, Aug.2002.

[8] T. Fawcett, and F. Provost, "Adaptive fraud detection," *Data Mining and Knowledge Discovery*, vol.1 (3), pp.291-316, Sep.1997.

[9] C.X. Ling, and C. Li, Applying Data Mining to Direct Marketing. In W. Kou and Y. Yesha (eds.), *Electronic Commerce Technology Trends: Challenges and Opportunities*, pp.185-198, IBM Press.2000.

[10] G M. Weiss and B D. Davison , *Data Mining in H. Bidgoli (Ed.), The Handbook of Technology Management*, Vol. II, John Wiley and Sons, 2010.

[11] P.N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, 2nd ed., Addison Wesley, 2006.

[12] S Chaudhuri and U Dayal, "An Overview of Data Warehousing and OLAP Technology," *ACM SIGMOD Record*, vol. 26(1), pp. 65-74, Mar.1997.