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

Ms. Jaspreet Kaur Gill

Assistant Professor in Computer Science & Application

S.D.S. College For Women, Lopon (Moga), India

Abstract: The world is deluged with various kinds of data-scientific data, environmental data, financial data and mathematical data. Manually analyzing and summarizing the data is impossible because of the incredible increase in data. Data mining is proved to be beneficial for making decisions by elicitation knowledge from heterogeneous sources. The focus of the paper is to review data mining architecture, tasks, techniques that user can use according to their requirements.

Keywords: Data, data mining, database, process, pattern, customer.

I. INTRODUCTION

Data mining is the process of turning raw data into useful information. Any numbers, text, facts, web pages or documents that can be processed by a computer are considered data and mining is the process of extracting something useful. Hence, as the name indicates, data mining is the process of extracting knowledge from large volumes of data. Data mining means digging through large volumes of data and extracting previously unidentified and potentially useful information. In other words, data mining comes up with information that queries or reports cannot discover normally. By finding out useful patterns and trends about different aspects of the company, businesses can come up with new strategies that are helpful in gaining competitive advantage.

II. DATA MINING ARCHITECTURE

There are number of components involved in data mining process. Each element encapsulates some specific functionality and linked with each other for solving data mining tasks. These modules constitute the architecture of data mining. The major components of any data mining system are data source, data warehouse server, data mining engine, pattern evaluation module, graphical user interface and knowledge base. These components are shown in Fig.1.

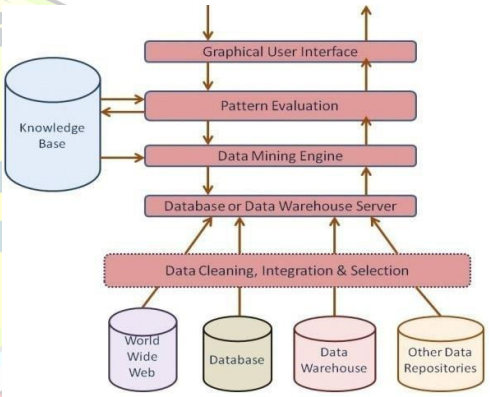

Fig.1 Components of Data Mining

A. *Data Sources*

Database, data warehouse, World Wide Web (WWW), text files and other documents are the actual sources of data. You need large volumes of historical data for data mining to be successful. World Wide Web or the Internet is another big source of data.

B. *Different Processes*

The data needs to be cleaned, integrated and selected before passing it to the database or data warehouse server. As the data is from different sources and in different formats, it cannot be used directly for the data mining process because the data might not be complete and reliable. So, first data needs to be cleaned and integrated.

C. *Database or Data Warehouse Server*

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

It contains the actual data that is ready to be processed. Hence, the server is responsible for retrieving the relevant data based on the data mining request of the user.

D. *Data Mining Engine*

It is the core component of any data mining system. It consists of a number of modules for performing data mining tasks including association, classification, characterization, clustering, prediction, time-series analysis etc.

E. *Pattern Evaluation Modules*

It is mainly responsible for the measure of interestingness of the pattern by using a threshold value. It interacts with the data mining engine to focus the search towards interesting patterns.

F. Graphical User Interface

It communicates between the user and the data mining system. This module helps the user use the system easily and efficiently without knowing the real complexity behind the process. When the user specifies a query or a task, this module interacts with the data mining system and displays the result in an easily understandable manner.

G. *Knowledge Base*

It is helpful in the whole data mining process. It might be useful for guiding the search or evaluating the interestingness of the result patterns. It might even contain user beliefs and data from user experiences that can be useful in the process of data mining. The data mining engine might get inputs from the knowledge base to make the result more accurate and reliable. The pattern evaluation module interacts with the knowledge base on a regular basis to get inputs and also to update it.

III. DATA MINING TASKS

The data mining tasks can be classified two categories are descriptive tasks and predictive tasks. Predictive data mining tasks come up with a model from the available data set that is helpful in predicting unknown or future values of another data set of interest. A medical practitioner trying to diagnose a disease based on the medical test results of a patient can be considered as a predictive data mining task. Descriptive data mining tasks usually finds data describing patterns and comes up with new, significant information from the available data set. A retailer trying to identify products that are purchased together can be considered as a descriptive data mining task. Divisions of these tasks are shown in Fig.2.

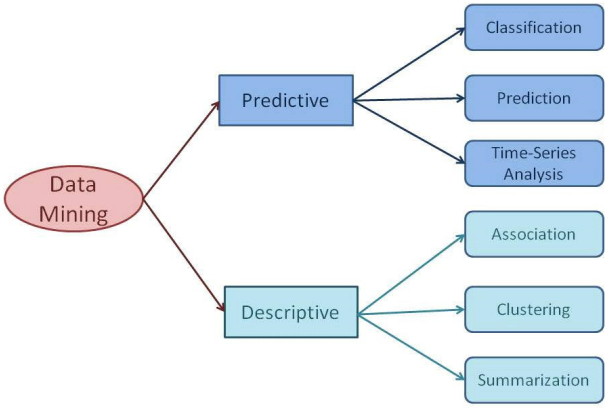

Fig. 2 Tasks of Data Mining

A. Classification

It derives a model to determine the class of an object based on its attributes. A collection of records will be available, each record with a set of attributes. One of the attributes will be class attribute and the goal of classification task is assigning a class attribute to new set of records. Classification can be used in direct marketing, that is to reduce marketing costs by targeting a set of customers who are likely to buy a new product. Hence, {purchase, don't purchase} decision forms the class attribute in this case.

B. *Prediction*

This involves developing a model based on the available data and this model is used in predicting future values of a new data set of interest. For example, a model can predict the income of an employee based on education, experience and other demographic factors like place of stay, gender etc.

C. Time - Series Analysis

This is a sequence of events where the next event is determined by one or more of the preceding events. Stock

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

market prediction is an important application of time- series analysis.

D. Association

Helps to discovers the association or connection among a set of items. A retailer can identify the products that normally customers purchase together or even find the customers who respond to the promotion of same kind of products. If a retailer finds that beer and nappy are bought together mostly, he can put nappies on sale to promote the sale of beer.

E. *Clustering*

This is used to identify data objects that are similar to one another. The similarity can be decided based on a number of factors like purchase behavior, responsiveness to certain actions, geographical locations and so on. For example, an insurance company can cluster its customers based on age, residence, income etc.

F. *Summarization*

It is the generalization of data. A set of relevant data is summarized which result in a smaller set that gives aggregated information of the data. For example, the shopping done by a customer can be summarized into total products, total spending, offers used, etc.

IV. DATA MINING TECHNIQUES

Data mining integrates approaches and techniques from various disciplines such as machine learning, association rule learning, neural networks, data visualization etc. In short, data mining is a multi-disciplinary field.

A. Machine Learning

Machine learning is the collection of methods, principles and algorithms that enables learning and prediction on the basis of past data. Machine learning is used to build new models and to search for a best model matching the test data. Data mining uses a number of machine learning methods including inductive concept learning, conceptual clustering and decision tree induction. A decision tree is a classification tree that decides the class of an object by following the path from the root to a leaf node. Fig. 3 is a simple decision tree that is used for weather forecasting.

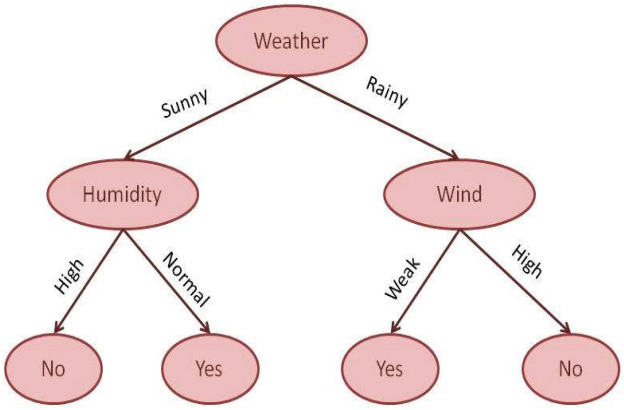

Fig.3 Example of decision tree

B. *Association rule learning*

It is a method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness.

Market Basket Analysis is one of the most common and useful types of data analysis for marketing and retailing. The purpose of market basket analysis is to determine what products customers purchase together. It takes its name from the idea of customers throwing all their purchases into a shopping cart (a "market basket") during grocery shopping. Knowing what products people purchase as a group can be very helpful to a retailer or to any other company. A store could use this information to place products frequently sold together into the same area, while a catalog or World Wide Web merchant could use it to determine the layout of their catalog and order form.

C. Neural network

It contains collections of connected nodes with input, output, and processing at each node. Between the visible input and output layers may be a number of hidden processing layers. Each processing unit (circle) in one layer is connected to each processing unit in the next layer by a weighted value, expressing the strength of the relationship. This approach is an

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

attempt to copy the way the human brain works· in recognizing patterns.

D. *Data Visualization*

The information get from large volumes of data should be presented well to the end user and data visualization techniques make this possible. Data is transformed into different visual objects such as dots, lines, shapes etc and displayed in a two or three dimensional space. Data visualization is an effective way to identify trends, patterns, correlations and outliers from large amounts of data.

V. APPLICATIONS OF DATA MINING

A. *Data Mining in Retail Industry*

The strength of data mining is effectively used mainly by customer focused businesses. Retailers can satisfy the customers by introducing price discounts or purchase coupons. Data mining can be used to identify consumer demand for the products and also to understand how the price change of a particular product affects sales of other products. Data mining techniques can be used to identify how effective a particular promotion could be across different media or geographic locations. For example, data mining can be applied to check which segment of customers respond positively to a promotion, which media channels have been successful for different campaigns in the past and so on. By analyzing this kind of information, retailer can come up with more effective and fruitful promotions and advertisements.

B. *Telecommunication Industry*

This is one of the rapidly growing and highly competitive businesses today. Telecommunication industry offers a number of services including cell phone, fax, internet messenger, web data transmission and so on. Lots of fraudulent activities are happening around misusing the potential of Internet and cell phone services. Data mining is really helpful in identifying such fraudulent activities.

C. *Financial Organizations*

Nowadays a number of banks and financial institutions offer a wide variety of banking services including investment, loans, credit cards etc. The data collected by these organizations are usually reliable, complete and of high quality. Hence, data mining can provide reliable information. Data mining can be effectively used to predict loan payment, identify fraudulent activities, analyze customer credit policy, classify and cluster customers for target marketing and so on.

VI. CHALLANGES IN DATA MINING

Though data mining is very powerful, it faces many challenges during its implementation. The challenges could be related to performance, data, methods and techniques used etc. The data mining process becomes successful when the challenges or issues are identified correctly and sorted out properly.

A. *Noisy and Incomplete Data*

Data mining is the process of extracting information from large volumes of data. The real-world data is heterogeneous, incomplete and noisy. Data in large quantities normally will be inaccurate or unreliable. These problems could be due to errors of the instruments that measure the data or because of human errors. Suppose a person might make spelling mistakes while entering the email id which results in incorrect data. Even some customers might not be ready to disclose their email id which results in incomplete data. The data even could get altered due to system or human errors. All these result in noisy and incomplete data which makes the data mining really challenging.

B. Distributed Data

Real world data is usually stored on different platforms in distributed computing environments. It could be in databases, individual systems, or even on the Internet. It is practically very difficult to bring all the data to a centralized data repository mainly due to organizational and technical reasons.

C. Complex Data

Real world data is really heterogeneous and it could be multimedia data including images, audio and video, complex data, temporal data, spatial data, time series, natural language

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

text and so on. It is really difficult to handle these different kinds of data and extract required information.

D. Data Visualization

Data visualization is a very importance process in data mining because it is the main process that displays the output in a presentable manner to the user. The information extracted should convey the exact meaning of what it actually intends to convey. But many times, it is really difficult to represent the information in an accurate and easy-to-understand way to the end user.

E. *Data Privacy and Security*

Data mining normally leads to serious issues in terms of data security, privacy and governance. For example, when a retailer analyzes the purchase details, it reveals information about buying habits and preferences of customers without their permission.

VII. SCOPE OF DATA MINING

Data mining draw valuable business information in large database for example, finding linked products in gigabytes or terabytes of store scanner data. Given database of sufficient size and good quality, data mining technology can generate new decisions making business opportunities by providing these capabilities. It automates the process of finding predictive information from large database. It uses current or past promotional mailings data to identify the most likely to maximize the return on investment.

VIII. CONCLUSION

In this paper, we briefly reviewed the various data mining concepts, its techniques and applications. It is not a new term, but in recent years its growth day by day touches great horizons. It has spread almost nowadays. This review would be helpful for the researchers to focus on the various issues of data mining. From above study it seems very difficult to design and develop a data mining system, which can work dynamically for any domain.

ACKNOWLEDGEMENT

The completion of this research paper could not have been possible without the expertise of Ms. Gurpreet kaur as she gave her active guidance throughout the completion of paper. Last but not the least, I would also want to extend my appreciation to those who could not be mentioned here but well played their role to inspire the curtain.

REFERENCES

[1] Arun K Pnjari : *Data MiningTechniques*,2nd Edition.

[2] Jiawei Han and MichelineKamber: *Data Mining Concepts and Techniques,* 2nd Edition, 2011.

[3] Pang-Ning Tan and Vipin Kumar: Intriduction to Data Mining, 2006.

[4] http://ecomputernotes.com/database-system/adv-database/data-mining

[5] http://www.tutorialpoint.com/data_mining/dm_quick_guide.html

[6] http://www.ijcse.com/docs

[7] http://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/

[8] http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_exports

[9] https://www.slideshare.net/mobile/macjones25/decision-tree-10164318

[10] https://web.fhnw.ch/personenseiten/taoufik.nouri/Data%20Mining/Course/Case%20Study/PA

[11] https://en.m.wikipedia.org/wiki/Association_rule_learning

BIOGRAPHY

Ms. Jaspreet Kaur, Msc(IT), MCA, Department of Computer Science & Application, S.D.S. College For Women, Lopon (Moga), India.