International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 5 Issue: 6 746 – 752

Challenging Problems in Data Mining and Data Warehousing

Esha Rana (MCA Student)
MCA, Mumbai University/ Vivekanand Education Society's Institute of Technology, Chembur
Mumbai, India
e-mail: esha.rana@ves.ac.in

Dashrath Mane (Professor)
MCA, Mumbai University/ Vivekanand Education Society's Institute of Technology, Chembur
Mumbai, India
e-mail: dashrath.mane@ves.ac.in

***Abstract*—** Data mining is a process which is used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers and develop more effective marketing strategies as well as increase sales and decrease costs. It depends on constructive data collection and warehousing as well as computer processing. Data mining used to analyze patterns and relationships in data based on what users request. For example, data mining software can be used to create classes of information. When companies centralize their data into one database or program, it is known as data warehousing. Accompanied a data warehouse, an organization may spin off segments of the data for particular users and utilize. While, in other cases, analysts may begin with the type of data they want and create a data warehouse based on those specs. Regardless of how businesses and other entities systemize their data, they use it to support management's decision-making processes.

Keywords-*Datawarehousing, Datamining*

*****

I. INTRODUCTION

Data mining can be defined as the extraction of hidden predictive information from large databases. Data Mining is a robust technology with great prospective to help companies concentrating on the important facts in their data warehouses. Data mining tools can anticipate future trends as well as etiquette allowing businesses to make knowledge-driven decisions. Data mining tools can answer business questions that traditionally were time consuming to resolve.Data warehousing means electronically storing massive amount of information by a business. Warehoused data should be stored in a manner that is secure, easy to retrieve and easy to manage. The notion of data warehousing derive in 1988 with the work of IBM researchers Barry Devlin and Paul Murphy. The need to warehouse data evolved as computer systems became more complex and can handle large amounts of data. If you purchase for example a gadget online, the website will recommend you products that you may want to get based on your preference. Also have you ever thought about the alerts you get from your bank when you use your credit card in a different city. These are examples of data mining which is the process of discovering useful patterns from a huge data set. This large data is created by integrating present and historical data from different sources and store them in a special repository called Data Warehousing (DW). DW is a very important repository especially for the historical data and non-every-day transactions. For example, historic data about the purchase transactions made by customers at a supermarkets. Keeping such type of data in a regular database will make it very huge and then slower performance. The ways of designing data warehousing and regular databases are different. Data warehousing design depends on a dimensional modeling technique. The multidimensional modeling (e.g. star schema) provides faster performance. Data Mining is used to provide useful information to technical and non-technical users which will help them to make better decisions. Data mining is a process used by companies to convert raw data into useful information. It is normally used as a decision support system. Data Mining is not an easy process. Sometimes the whole process needs to be repeated. For this reason the data mining process is considered as an iterative process. It consists of six phases: 1) problem definition, 2) data preparation, 3) data exploration 4) modeling, 5) evaluation, and 6) deployment. Data Mining automates the process of extracting information. This is the reason why it is used in different areas, especially business where it is important to analyze huge amount of data. One of the most common uses of data mining is web mining. The Internet has become important part of our life. Terabytes of data being added every day, extracting the information using the data mining techniques has become very important.

II. DATA MINING MODELS

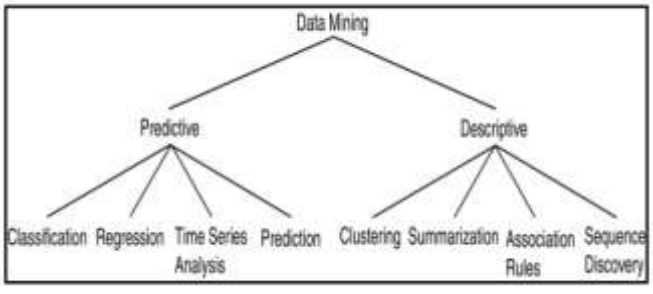

Figure 1 Data Mining Models

Data mining is also called knowledge Discovery in databases. Data mining is a process of exploration an analysis, by semi automatic means of large quantities of data in order to discover meaningful models and rules.

Data mining means collection of information from unstructured data. So it is able to help achieve specific objectives. The purpose of a data mining effort is normally either to create a descriptive model or a predictive model A descriptive model presents, the main characteristics of the data set. The goal of a predictive model is to allow the data miner to predict an unknown value of a specific variable; the target variable. The goal of predictive and descriptive model can be achieved using a variety of data mining.

Classification: Classification is based on categorical. This technique based on the supervised learning. This can be classifying the data based on the training set and values (class label). These goals are achieve using a decision tree, neural network or classification rule.

for example we can apply the classification rule on the past record of the student who left for university and evaluate them. Using these techniques we can easily identify the performance of the student.

Regression: Regression is used to map a data item to a real valued prediction variable. In other words, regression can be adapted for prediction. In the regression techniques target value are known. For example, you can predict the child behavior based on family history.

Time Series Analysis: Time series analysis is the process where one uses of statistical techniques to model a time-dependent series of data points. Time series forecasting is a method of using a model to generate forecast for future events based on known past events.. For example stock market.

Prediction: It is a data mining method that finds the relationship between independent variables and also the relationship between dependent and independent variables. Prediction model is based on continuous or ordered value.

Clustering: Clustering is a collection of similar data object. Dissimilar object form another cluster. It is way finding similarities between data according to their characteristic. This technique based on unsupervised learning. For example, image processing, pattern recognition, city planning.

Summarization: Summarization is abstraction of data. It is set of relevant task and gives an overview of data.

Association Rule: Association is the most popular data mining techniques and fined most frequent item set. Association strives in discovering patterns in data which are based upon relationships between items in the same transaction. Because of this, association it is often referred to as "relation technique". This technique of data mining is used within the market based analysis in order to identify sets of goods that consumers often purchase at the same time.

Sequence Discovery: A technique of uncovering relationship among data. It is set of object each associated with its own timeline of events. For example, scientific experiment, natural disaster and analysis of DNA sequence.

III. DATA MINING APPLICATIONS IN REAL WORLD

A number of fields have adapted data mining technologies because of fast access of data and valuable information from a large amount of data.. Some of the main applications listed below:

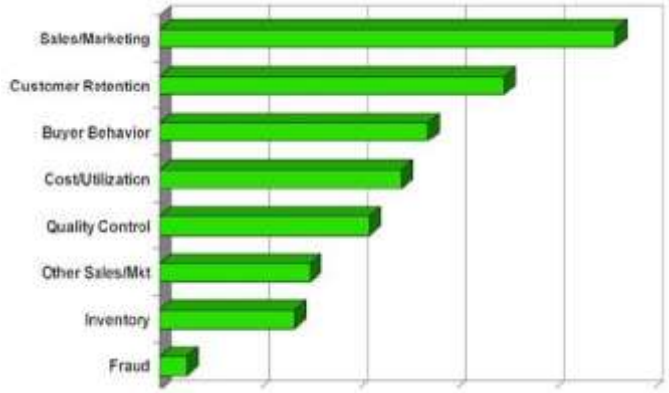

Figure 2 Data Mining Applications

Data Mining in Education Sector: We apply data mining in education sector. Using these term enhances the performance of student, drop out student, student behavior, which subject selected in the course. Using student's data to analyze their learning behavior to predict the results.

There is a growing field, known as Educational Data Mining, it deals with with developing methods that discover knowledge from data originating from educational Environments. The goals of EDM are identified as predicting student's future learning behavior, studying the effects of educational support, and advancing scientific knowledge about learning. Data mining can be used by an institution to take more accurate decisions and to predict the results of the student. With the help results they can focus on what to teach and how to teach. Learning pattern of the students can be captured and used to create methods to teach them.

Data Mining in Banking and Finance: Data mining has been pre-owned predominantly in the banking and financial markets. In the banking field, data mining is used to predict credit card fraud and in estimating risk as well as to examine the trend. In the financial markets, data mining approach such as neural networks is used in stock forecasting, price prediction.

Data Mining in Market Basket Analysis: These methods are based on shopping database. The ultimate goal of market basket analysis is finding the products that customers quite frequently purchase. The stores then use this information by putting these products in close proximity of each other and making them more visible for customers at the time of shopping.

Data Mining in Earthquake Prediction: Predicting earthquakes from the satellite maps. There are two basic classifications of earthquake predictions: forecasting (months to years in advance) and short-term predictions (hours or days in advance).

Data Mining in Bioinformatics: Bioinformatics generate a large amount of biological data. The importance of this new field of inquiry will grow as we continue to generate and integrate large quantities of genomic, proteomic, and other data.

DataMining in Telecommunication:The telecommunications fields have made use of data mining technology because telecommunication industry has huge amounts of data and have a very large number of customer, and rapidly changing and highly competitive environment. Telecommunication companies use data mining technique to improve their marketing, detection of fraud, and better management of telecommunication networks.

Data Mining in Agriculture: Data mining has been emerging in agriculture field for crop yield analysis. The yield prediction problem can be solved by deploying Data Mining techniques such as K Means, K nearest neighbor (KNN), Artificial Neural Network and support vector machine (SVM). .District wise agricultural production area and yield of crops is compiled, analysis, mining and forecasting. Statistics on consumption of fertilizers can be converted into a data merge. Data on agricultural inputs like seeds and fertilizers can also be effectively analyzed in a data warehouse. Data from livestock census can be converted into a data warehouse. Land use pattern statistics can also be analyzed in a warehousing environment.

Therefore, there is great scope for application of data warehouse housing and data mining techniques in agricultural sector.

Data Mining in Cloud Computing: Data Mining techniques are used in cloud computing. The implementation of data mining techniques through Cloud computing will allow the users to retrieve information from virtually integrated data warehouse that reduces the costs of infrastructure and storage Cloud computing uses the Internet services that rely on clouds of servers to handle tasks. The data mining technique in Cloud Computing to perform efficient, reliable and secure services for their users.

CRM: Customer Relationship Management is all acquiring and retaining customers. To maintain a proper relationship with a customer a business need to collect data and analyze the information. This is where data mining plays its role. With data mining technologies the collected data can be used for analyzing information. Instead of being confused where to focus to retain customer, the seekers for the solution get filtered solutions.

Corporate Surveillance: Corporate surveillance is the monitoring of a single user's or group's behavior by a corporation. The data collected is most frequently used for marketing grounds or sold to corporations, but is also habitually shared with government. It can be used by the business to tailor their products desirable by their consumers. The data can be used for direct marketing purposes, such as the targeted advertisements on Google and Yahoo, where ads are targeted to the user of the search engine by analyzing their search history and emails.

IV. DATAWARE HOUSING APPLICATIONS IN REAL WORLD

Enumerated below are the applications of Data warehouses .

We are going to discuss various applications of data warehouse:

Figure 3 Data Warehouse Applications

Banking Industry: In the banking industry, priority is given to risk management and policy reversal and also for analysing consumer data, market trends, government proclamation importantly financial decision making. Most banks also use warehouses to manage the resources available on deck in an productive manner. undisputed banking sectors deploy them for market research, performance analysis to develop marketing programs.

Analyzing card holder's transactions, spending patterns which provide the bank with an opportunity to introduce special offers and deals based on cardholder activity.

Consumer Goods Industry: These are used for predicting consumer trends, inventory management, market research. In-depth analysis of sales and production are carried out. Apart from these, information is exchanged business partners.

Government and Education: The government can also use it for services related to human resources like recruitment, and accounting like payroll management. The government uses data warehouses to maintain and analyze tax records, health policy records and also entire criminal law database is connected to the data warehouse. Actions of criminals can be predicted from the patterns the analysis of historical data related with past criminals .Universities make use warehouses for extracting of information used for the proposal of research grants, understanding their student demographics, and human resource management.

Healthcare: One of the most important sectors which utilizes data warehouses is the Healthcare sector. All of financial, clinical, and employee records are put into warehouses as it helps them to predict outcomes, track and analyze their service feedback, generate patient reports, share data with tie-in insurance companies, medical aid services, etc.

Hospitality Industry: A major proportion of this industry is dominated by hotel and restaurant services, car rental services, and holiday home services. They deploy warehouse services to design and estimate their advertising and promotion campaigns where they aim customers based on their feedback and travel patterns.

Insurance: The warehouses are primarily used to analyze data patterns and customer trends, apart from maintaining records of already existing participants.

Manufacturing and Distribution Industry: A manufacturing organization has to take several make-or-buy decisions which can influence the future of the sector, which is why they utilize high-end OLAP tools as a part of data warehouses to estimate market changes, analyze current business trends, identify warning conditions, view marketing developments, and ultimately take better conclusion. They additionally use them for product shipment records, records of product portfolios, identify profitable product lines, analyze previous data and customer feedback to evaluate the weaker product lines and eliminate them. For the distributions, the supply chain management of products operates through data warehouses.

The Retailers: Retailers are middlemen between producers and consumers. It is important for them to maintain records to ensure their existence in the market. They use warehouses to track items, advertise promotions, and the consumers buying trends. They also examine sales to direct fast selling and slow selling product lines and examine their shelf space through a process of elimination.

Services Sector: Data warehouses find themselves to be of use in the service sector for maintaining financial records, revenue patterns, profiling of customer , resource management, and human resources.

Telephone Industry: The telephone industry controls over both offline and online data burdening them with a lot of historical data which has to be strengthen and united. Apart from those operations, analysis of fixed assets, analysis of customer's calling patterns for sales representatives to push advertising campaigns, and tracking of customer queries, all require the facilities of a data warehouse.

Transportation Industry: In the transportation industry, data warehouses record customer data enabling traders to testing with target marketing where the marketing campaigns are designed by keeping customer requirements in mind. The internal environment of the industry uses them to examine customer feedback, performance, organize crews on board as well as examine customer financial reports for pricing strategies.

Marketing: Every business is not successful without proper marketing and marketing. Marketing requires knowing the latest trends and demands. DW in marketing is used to examine the patterns of customer's behavior and use this customer information for implementing relationship marketing. They play a vital role in identifying and targeting the customers

a) Trend Analysis it is a technique that is used to predict future outcomes from historical information. Divergent medium to large scale enterprises are following this. In trend analysis, Data Warehouse can be used to analyse behaviors of the consumers by using historical records over months.

b) Web Marketing Web is a hub of millions of devices .It refers to a category of advertising that includes any marketing activity conducted online. Facebook, Google, uses web marketing and are depending on latest updated data DW.

c) Market Segmentation Behavior identification is priority of any organization. Market segmentation is the identifying the customer's behavior and common characteristics related to the purchases made against that product of related company.

V. CHALLENGING PROBLEMS IN DATA MINING

Developing a Unifying Theory of Data Mining: Several respondents feel that the current state of the art of data mining research is too "ad-hoc." Many techniques are designed for individual problems, such as classification or clustering, but there is no unifying theory. However, a theoretical framework that unifies different data mining tasks including clustering, classification, association rules, etc., as well as different data mining approaches (such as statistics, machine learning, database systems, etc.), would help the field and provide a

basis for future research. There is also an opportunity and need for data mining researchers to solve some longstanding problems in statistical research, such as the age-old problem of avoiding spurious correlations.

Scaling Up for High Dimensional Data and High Speed Data Streams: One challenge is how to design classifiers to handle ultra-high dimensional classification problems. There is a strong need now to build useful classifiers with hundreds of millions or billions of features, for applications such as text mining and drug safety analysis. Such problems frequently begin with tens of thousands of features and also with interactions between the features, so the number of involved features gets huge quickly. One important problem is mining data streams in extremely large databases (e.g. 100 TB). Satellite and computer network data can easily be of this scale.

Although, now a days data mining technology is still too slow to handle data of this scale. Additionally, data mining should be a continuous, online process, rather than an occasional one-shot process. Organizations that can do this will have a significant advantage over ones that do not. Data streams used to present a new challenge for data mining researchers. One particular instance is from high speed network traffic where one hopes to mine information for various purposes, including identifying anomalous events possibly indicating attacks of one kind or another. A technical problem is how to compute models over streaming data, which adapt changing environments from which the data are drawn.

Mining Sequence Data and Time Series Data: Sequential and time series data mining is an important problem. Despite progress in other related fields, how to efficiently cluster, classify and predict the trends of these data is still an important open topic.

A specially challenging problem is the noise in time series data. It is an important open issue to tackle. Many time series used for predictions are contaminated by noise, making it difficult to do accurate short-term and long-term predictions.

Examples of these applications include the predictions of financial time series and seismic time series. Although signal processing techniques, such as wavelet analysis and filtering, can be applied to remove the noise, they often introduce lags in the filtered data. Such lags reduce the accuracy of predictions because the predictor must overcome the lags before it can predict into the future. Existing data mining methods also have difficulty in handling noisy data and learning meaningful information from the data. Some of the key issues that need to be addressed in the design of a practical data miner for noisy time series include:

• Information/search agents to get information: Use of wrong, too many, or too little search criteria; possibly erratic information from many sources; semantic analysis of (meta-) information; assimilation of information into inputs to predictor agents.

• Learner/miner to modify information selection criteria: allocating of biases to feedback; developing rules for Search Agents to collect information; developing rules for Information Agents to assimilate information.

• Predictor agents to predict trends: Incorporation of qualitative information; multi objective optimization not in closed form.

VI. CHALLENGING PROBLEMS OD DATA WAREHOUSING

Nevertheless of going through huge amount research during the last decade Data warehouse still have several areas to research and improve. Some of the major issues to be tackled are as follows:

Data extraction and cleaning are the first step to construct a data warehouse. For any kind of database the quality of data is the most important characteristics to get the desired output efficiently. Today we have number of tools available for Data extraction and Cleaning but they are not providing the desired productivity.

The source system consists of all those 'Production' raw data providers, from where the details are pulled out for making it satisfactory for Data Warehousing. All these data sources are having their own techniques of storing data. Some of the data sources are cooperative and some might be non cooperative sources. Because of this diversity various reasons are present which may contribute to data quality problems, if not properly taken care of. A source that offers any kind of unsecured access can become unreliable-and ultimately contributing to poor data quality. Different data Sources have different kind of problems connected with it such as data from legacy data sources do not even have metadata that describe them. The sources of dirty data include data entry error by a human or computer system, data update error by a human or computer system. Part of the data comes from text files.

Data transformation and integration is an additional area to be researched further as data warehouse is constructed using data from heterogeneous sources hence we should have efficient tools then available at present. This is one of the most important tasks in data warehousing as different databases have different schemas and format and it's essential to convert them to similar format before loading into the data warehouse. The transformation of data with least error and least loss of information is still to go miles ahead.

Maintenance of a data warehouse is additional aspect in which we have lot of chances to improve. We should look for some better maintenance technologies along with the software and

better hardware to efficiently manage the increasing size of the data warehouse.

CONCLUSION

Data Mining and Warehousing talks about the modification in business trends these days. All the small and big industries are collecting and using data from various sources to identify their own business trends. Organizations understand the strengths and the weaknesses of their competitor improve their progressing speed towards the goal and expand their business empire. A data warehouse is a solution to a business problem not a technical problem. The data warehousing and data mining need to persistently overcome hurdles that are yet undefined and help the organization in decision making and improves the goodwill of organization.

Data mining helps in securing and processing the data into understandable chunks, where warehousing helps in analysing the data and put it in such a way as to facilitate comparison between trends, analyzing the data for the business predictions and accelerate decision making. In short, a data warehousing and data mining implementation includes the conversion of data from various source systems into a common format with accuracy, help the organization in the strong business resolution and help to expand the business empire. A Data Warehouse Enhances Consistency and Data Quality each data from the various departments is standardized, each department will produce results that are in line with all the other departments. It is applicable and organized in an efficient manner. One strong feature of data warehouses is that data from different locations can be combined in one location.

Data warehousing provides the means to change the raw data into information for making effective business decisions which emphasis on information, not data. The data warehouse is the hub for decision support data. Data mining is a useful tool with multiple algorithms that can be tuned for specific asks. It can benefit business, medicine, and science. It needs more efficient algorithms to speed up data mining process.

Data Mining brings a lot of benefits to businesses, society, governments as well as individual. However privacy, security and misuse of information are the major problems if they are not addressed and resolved properly.

Moreover, since the data mining process is systematic, it offers enterprises/government the ability to discover hidden patterns in their data-patterns that can help them understand customer behaviour and market trends. In this paper, the concept of data mining, role of data mining its major challenges, issues and application have been focused which help in business strategy formulations, decision making and analysis to the business, society and governments.

ACKNOWLEDGMENT

The author would like to thank the management and prof. Dashrath Mane , Vivekanand Education Society's Institute of Technology for the guidance and suggestions throughout the project.

REFERENCES

[1] Stolba, N., Banek, M. and Tjoa, A.M. (2006): The Security Issue of Federated Data Warehouses in the Area of Evidence- Based Medicine. Proc. of the First
[2] International Conference on Availability, Reliability and Security (ARES'06,IEEE), 20-22 April, 2006.
[3] Inmon, W. (2002): Building the Data Warehouse, 3rd edition, Wiley-NewYork.
[4] SAS© (2002): Building a Data Warehouse Using SAS/Warehouse Administrator®, Software Course Notes (Book code58787). SAS Institute Inc.,Cary, NC 27513, USA.
[5] Multidimensional Data analysis and Data Mining- Arinjay Choudhary, Dr. P.S. Deshande
[6] Data Mining and Data Warehousing and OLAP-A. Berson, S.J. Smith
[7] J. Hammer, H. Garcia-Molina, J. Widom, W. Labio, and Y. Zhuge. The Stanford Data Warehousing Project. IEEE Data Engineering Bulletin, Special Issue on Materialized Views and Data Ware housing, 18(2):41{48, June 1995.
[8] W.H. Inmon and C. Kelley. Rdb/VMS: Developing the Data Warehouse. QED Publishing Group, Boston, Massachussetts, 1993.
[9] A. Gupta and I.S. Mumick. Maintenance of materialized views: Problems, techniques, and applications. IEEE Data Engineering Bulletin, Special Issue on Materialized Views and Data Warehousing, 18(2):3{18, June 1995}.
[10] Providing Architecture of the Data warehouse.http://it.toolbox.com/wiki/index.php/Data_warehouse_Architecture. Providing ETL Back end tools. [http://etltool.com]
[11] Jiawei Han, Micheline kamber, Jian pei "Data mining Concepts and Techniques" Third edition.
[12] QIANG YANG Department of Computer Science Hong Kong University of Science and Technology Clearwater Bay, Kowloon, Hong Kong, China XINDONG WU Department of Computer Science University of Vermont 33 Colchester Avenue, Burlington, Vermont 05405, USA xwu@cs.uvm.edu
[13] Milija SUKNOVIĆ, Milutin ČUPIĆ, Milan MARTIĆ Faculty of Organizational Sciences, University of Belgrade, Belgrade, Serbia and Montenegro Milijas@fon.bg.ac.yu, Cupic@fon.bg.ac.yu, Milan@fon.bg.ac.yu Darko KRULJ Trizon Group, Belgrade, Serbia and Montenegro KruljD@trizongroup.co.yu
[14] Lei Hu ,School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China
[15] Le Huifeng, Lin Jiajun, "Data mining technology and its application in process control [J]", *World instruments and automation*, vol. 3, no. 4, pp. 22-25, 1999.

ISSN: 2321-8169
746 – 752

[16] Ming Fan, Meng Xiaofeng, Data mining: concepts and techniques [M], Beijing:Machinery Industry Press, pp. 1-322, 2001.
[17] Jennifer Widom Department of Computer Science Stanford University Stanford, CA 94305-2140 widom@db.stanford.edu
[18] Mr. Dishek Mankad1 , Mr. Preyash Dholakia2 1 M.C.A., B.R.Patel Institute of Computer Application [MCA Program] 2M.C.A. K.S.K.V. Kachchh University MCA College
[19] Kimball, R. The Data Warehouse Toolkit. John Wiley, 1996.
[20] Barclay, T., R. Barnes, J. Gray, P. Sundaresan, "Loading Databases using Dataflow Parallelism." SIGMOD Record, Vol. 23, No. 4, Dec.1994.