International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 4, April - 2014

Data Mining Concepts, Technologies and Applications

Rashmi Bhatia
Dept. of Computer Applications
Dev Samaj College for Women
Chandigarh, India

***Abstract*— Data Warehousing and Data Mining are developing at a growing pace having a bright future. It has already been adopted by a large number of organizations and still scope is increasing day by day. No doubt that creating a data warehouse and then managing it is very hard but still more and more data warehouse systems and data mining tools are made available in the market. It is a hot topic of research and is being applicable in almost every field. But yes, business community, by using these technologies, is gaining competitive advantage by deriving insight and intelligence out of the data, so as to increase the revenue, reduce the cost and compete effectively; manage the complexity of the business by mining of data to unearth hidden patterns and trends; investing avenues in most effective way etc. In this paper, emphasis is put on Data Mining models, techniques, working, its relationship with Data Warehousing, and also applicability of Data Warehousing and Data Mining in real world.**

Keywords*— *knowledge Discovery in Databases (KDD); Inductive Learning; Cluster; STATISTICA

I. INTRODUCTION

"Data Mining can be described as exploratory data analysis. The aim is to look for intersection patterns in the data, patterns that can be used to set business strategy or to identify unusual behavior. Data mining tools apply statistical techniques to large quantities of stored data in order to look for such patterns". (Date, C.J., 2000).

Just like mining stands for mining a mountain for a vein of valuable ore, the data mining refers to computer-assisted process of digging through and analyzing voluminous data and finding out valuable information used to predict behaviors and future trends, allowing business to make pro-active, knowledge-driven decisions.

Data mining is also referred to as knowledge Discovery in Databases (KDD) because a variety of new knowledge can be discovered from the database after going through six phases: selection, preprocessing, transformation, data mining, reporting and display of discovered information**.**

II. RELATION BETWEEN DATA WAREHOUSE AND DATA MINING:

Data warehouse lays foundation of data mining by integrating data in proper format and data mining extracts meaningful new patterns from data warehouse, which would not have been possible merely through querying or processing of data.

- In early times, data used to be collected for the purpose of creating records and maintaining databases. DBMS was used in this context
- RDBMS replaced DBMS to provide a better access to the data, being more specific in making query
- Data warehouse provide a large repository of data, using multidimensional databases, so as to support decision-making
- Data mining helps in doing Business Trend Analysis and making prediction based on the data stored in data warehouse, using advanced computer algorithms.

Data mining development is based on a number of fields like:

A. *Inductive* Learning

"a device learns a concept if it can, given positive and negative examples, produce an algorithm that will classify future examples with some probability." (Rich, E. and Knight, K., 1991). This process of classification identifies classes in a way that each class has a unique pattern of values, which describes the class. It has two strategies

- Supervised Learning, where a teacher defines classes and gives examples of each class.
- Unsupervised Learning, where the system has to observe the examples and recognize the patterns by itself.

A. *Statistical* Reasoning

In this knowledge is represented through various statistical measures such as probability factors, certainty factors and rule-based systems, fuzzy logic and other techniques, but the results may be difficult to interpret. So data mining applies expert's knowledge of the data with advanced analysis techniques for better results.

B. *Machine* Learning

Machine learning allows the automation of learning process, through construction of rules based on observation of environmental states and transition. Machine learning includes data mining in respect of finding valuable knowledge in large set of real world examples.

III. DATA MINING MODELS

A. *Verification Model*

This model takes hypothesis formulated by the user and tests its validity against the data. But here user is held responsible for providing the model with the hypothesis and issuing the query on the data to affirm or negate the hypothesis. But in this model, the query always return records to verify or negate the hypothesis and no new information is created.

B. *Discovery Model*

This model automatically discovers the required valuable information hidden in the data in shortest possible response time. The data discovery of information is based on continuous observation without any user intervention and guidance.

IV. WORKING OF DATA MINING:

The methods adopted by data mining to perform various analyses are as follows:

A. *Classification*

"Classification is the process of assigning, to a particular input, the name of a class to which it belongs. The classes from which the classification procedure can choose can be described in a variety of ways. Their definition will depend on the use to which they will be put." (Rich, E. and Knight, K., 1991).

By this method a hierarchy of classes is created from an existing set of events or transactions. E.g., a population can be divided into different ranges of credit worthiness on the basis of the previous credit transactions.

B. *Association Rules*

It means associations between various items that people buy. For example, if someone buys bread, then most probably he would buy milk too.

The association rules consist of population, support and confidence.

- Population consists of a set of instances. E.g., Number of purchases made in the grocery store, each purchase represents an instance.
- Support is that fraction of population that satisfies both the antecedent and the consequent of the rule. e.g., out of all purchases only 2% include milk and potatoes
- Confidence is a measure of how often, if the antecedent is true, the consequent is also true. e.g., 90% of the purchases include both milk and bread.

C. *Sequential Patterns*

In this the behavior patterns and trends are anticipated. e.g., the customers repeating the purchase of the same item over a period of times can be known.

D. *Clusters*

A cluster is a set of objects grouped together on the basis of their logical relationships or consumer preferences. e.g., Data can be mined together to identify the various age groups of customers.

V. DATA MINING TECHNOLOGIES:

A. *Cluster Analysis*

Clustering is basically concerned with decomposing the database either according to similarity or according to optimization of set functions.

B. *Neural Networks*

Neural networks are non-linear predictive models with ability to learn. These have ability to derive meaning from complicated or even imprecise database and are used for extracting patterns and detecting trends, that are hardly noticed by other computer mechanisms.

C. *Genetic Algorithms*

These are the optimization techniques based on the processes such as genetic combination, mutation and natural selection.

D. *Decision Trees*

It is a tree-shaped structure used for knowledge representation where there are finite numbers of classes in which all the examples are classified.

E. *Nearest Neighbor Method*

A technique that classifies each record based in a dataset based on the records most similar to it in an historical dataset.

F. *Data Visualization*

It provides the analyst with a deeper, more intuitive understanding of the data by visual interpretation of complex relationships in multidimensional data, using various graphical tools, so that analyst's job is made easy and better predictions can be made.

VI. REAL WORLD APPLICATION OF DATA WAREHOUSE AND DATA MINING

The usage of the data warehousing and data mining technologies is increasing because of their applicability in various fields. There are certain fields in which the data warehousing and data mining are applicable, which are as follows:

A. *Banking*

- To store the large amount of data in data warehouse, even from the various branches
- To keep historic data, as any sort of information may be needed at any time to make future decisions
- To view the list of customers good or bad in their credit payments
- Number of credit card holders and their spending through credit card can be known

B. *Manufacturing*

- To optimize the use of men, machine and material
- To generate the list of frequency of machine failures
- To know the raw material used and will be used
- The manpower required for various tasks can be known

- To optimize the design of manufacturing process, shop-floor layouts and product design

C. Finance

- To analyze the credit worthiness of clients
- Evaluation of fraud detection
- To make analysis of proper canalizations of the funds so as to maximize the gain
- To know the future revenue expected, so that plans can be made to make the investments.

D. Marketing/Retail

- To analyze the customer behavior on the basis of the buying patterns
- Find associations between various products purchased
- To know the customers who can opt for the new products being launched, on the basis of their previous buying habits
- To plan a better advertising campaign
- Market basket analysis
- To design the store layout and catalogs

E. Health Care/Medicine

- To know the expected visits of customer to hospital
- To analyze the effective drugs for various diseases and know the side effects reported for various drugs
- To analyze the past diseases of a patient so as to relate it with present
- To identify successful medical therapies for different diseases

F. Library

- To analyze the demands of various books, so as to place an order of most demanded books first
- To have connectivity with other librarians too so as to see if a book is available there
- To analyze the trends in stealing of books

G. Transportation

- To decide the best route to be followed and schedule adopted
- To analyze the loading patterns

H. Insurance

- To categorize the customers on the basis of policies adopted
- To keep the previous details of the claim taken
- To have a check on fraudulent customers
- To decide upon the new policies based on the past trends

These were the various fields where the data warehousing and data mining is applicable but Data warehousing and data mining can be applied anywhere, where large amount of data storage is required, historical data is a need, lots of daily transactions take place and various analysis are required to be made.

VII. REAL WORLD EXAMPLES OF USAGE OF DATA WAREHOUSING AND DATA MINING:

A. Bass Brewers

Bass Brewers, the leading beer producers in U.K., opted for data warehousing and data mining. Rowley, N., the Information Infrastructure Manager states that for the first time people will be able to do data mining, ask questions we never dreamt, we could get the answers to, look for patterns among the data we could never recognize before.

The Bass Brewers are getting the benefits of data warehousing and data mining, by taking better decisions for growth of the organization.

B. Wal-Mart

Wal-Mart, USA's largest supermarket chain, having 2900 stores in 6 countries, established a data warehouse where all the point-of-sales transactions from over all the stores are continuously transmitted. They use data mining to process various queries. In 1995 on a record, they processed more than 1 million complex data queries.

C. Harvard University

Harvard University has developed a centrally operated fund-raising system; call HOLDEN (Harvard Online Development Network) that allows university institutions to share fund- raising information.

D. J.P.Morgan

J.P. Morgan, the leading financial company are using data mining through Information Harvester Software on the Convex Exampler and C series,.

E. Northern Bank

Northern Bank, a subsidiary of National Australia Group is also making use of a system, which integrates the multiple data sources into a consolidated database and then data mining is used to deliver that information to the user in a meaningful way

F. The National Basketball Association (NBA)

NBA applies data mining to analyze the movements of the players from the video clips stored, so as to help the coaches to make important selection decisions for the team.

G. Delphic Universities

Delphic Universities, a group of 24 universities, have opted for Holos for their management information system needs.

H. BeurBase, Amsterdam

BeurBase, Amsterdam, a real-time on-line Stock Exchange Relational Database receives data from ASE, Amsterdam Stock Exchange, stores it in data warehouse, contains real-time and historical data. Then the data is used by various applications for many research, education and public relation activities.

I. TSB Group PLC

TSB Group PLC is making use of Holos, supplied by Holistic Systems, which provide them a unique multidimensional functionality and flexibility.

J. AT & T, A.C. Nielson, and American Express

These are implementing data mining techniques for sales and marketing.

K. Merck-Medco Managed Care

Merck-Medco Managed Care, a mail-order business, basically dealing in selling drugs to various health care providers, has established its one terabyte data warehouse and is mining that data, so as to get better results in finding more effective treatment with lesser cost.

Other than above, there are certain data warehouse systems available in the market in large extent. One of the examples of such system is STATISTICA Data Warehousing System.

"STATISTICA Data Warehousing System is a complete powerful, scalable, and customizable intelligent data warehouse solution, which also optionally offers the most complete analytic functionality available on the market, fully integrated into the system". (http://www.statsoft.com/products/brochures/pdf/datawarehouse.pdf)

VIII. IMPORTANCE OF DATA WAREHOUSING AND DATA MINING TO DATABASE WORLD

- Dependence of business data warehouses is tremendously increasing because
  - → Data warehouse is the solid foundation for entire, enterprise wide business intelligence system. It is a system for intelligent management of unlimited amount of data, distributed across a large number of locations worldwide
  - → It provides more cost-effective decision-making by reduction of staff and computer resources needed to support queries.
  - → Multi-tiered data systems of data warehouse provide increased quality and flexibility of Business Trend Analysis.
  - → By correlating the customer's data in a data warehouse, better customer relationships can be established.

- Mento, B and Rapple, B. (2003, p.10) states that data mining has been integrated into the curriculum at many academic institutions. The scope of disciplines utilizing data mining is very broad, with the heaviest concentrations in the social sciences.

- No doubt that, data warehousing and data mining already are being adopted by a large number of organizations but still there are many interesting avenues for research in
  - → Data cleaning
  - → The physical design of data warehouse with respect to index selection, data partitioning and selection of materialized views
  - → The management of data warehouses with respect to detecting runaway queries, scheduling resources and managing them.

- Gatziu, S. and Vavouras, A. (1999, p.7) states that According to the market research firm Meta Group, the proportion of companies implementing data warehousing exploded from 10% in 1993 to 90% in 1994, and the data warehousing market will expand from $2 billion in 1995 to $8 billion in 1998.

- The future of data mining is prospering. If it is important and profitable to business concerns in short-term, then it would be very common to the public in middle-term, where it ca be used to find the best airfare to New York, finding contact number of a long-lost classmate etc. but in long-term, the possibilities are much advance. As data mining has its roots in Artificial Intelligence and statistics, it may result in Intelligent Agents.

- It has brought a turning point in the data base world, where traditional databases were developed to receive the information, were not suitable to support the decision making environment. The RDBMS are replaced by the multidimensional databases. The data warehousing and data mining implemented together make it possible to solve the queries or questions seem to be impossible. The impossibles are made possible and data base world has got a new face. If on the one hand, the size to the database has increased then on the other hand the complex queries are being processed to make the system more powerful.

Thus, the requirement of the age, the rich features, the design to support the future needs and the research taking place for these topics make them really advance and important. This is the reason that I have opted for this very topic.

CONCLUSION

From the above discussion, it is clear that, the whole community is leaning towards implementing these technologies. Also widespread availability of data mining tools provides a new avenue for various organizations to explore its potential in academic research and decision-making.

REFERENCES

[1] R. Elmasri and S.B. Navathe, Fundamentals of Database Systems, 3rd ed., Singapore: Pearson Education, 2000.
[2] C.J. Date, An Introduction to Database Systems, 7th ed., Singapore: Pearson Education, 2000.
[3] B. Mento and B. Rapple, Data Mining and Data Warehousing [online]. Washington, D.C: Association of Research Libraries, 2003. <http://www.inst-informatica.pt/servicos/informacao-e-documentacao/dossiers-tematicos/dossier-tematico-no-8-business-intelligence-abril-2010/artigos/data-mining-and-data-warehousing-mento-barbara-rapple-brendan>
[4] E. Rich and K. Knight, Artificial Intelligence, 2nd ed., New Delhi: Tata McGraw-Hill Publishing Company Limited, 1991.

[5] A. Silberschatz, H.F. Korth and S. Sudarshan, Database Systems Concepts, 4th ed., New York: McGraw-Hill, 2002.
[6] A. Vavouras and S. Gatziu, Data Warehousing: Concepts and Mechanisms [online]. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.88.1043&rep=rep1&type=pdf>
[7] S. Chaudhuri and U. Dayal, An Overview of Data Warehousing and OLAP Technologies [online]. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.151.8539&rep=rep1&type=pdf>
[8] Data Mining: What is Data Mining [online]. <http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm>
[9] D. Alexander, Data Mining [online]. <http://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat/Alex/>
[10] A. Perkins, Developing a Data Warehouse: The Enterprise Engineering Approach. [online]. <http://www.ies.aust.com/PDF-papers/dw.pdf>
[11] Statistica: Data Warehouse. [online]. <http://www.statsoft.com/Portals/0/Support/Download/Brochures/Data_Warehouse.pdf>