An Empirical Study on Data Mining Applications

P.Sundari[1] Dr.K.Thangadurai[2]

GJCST Computing Classification
H.2.8, J.1

***Abstract*-The wide availability of huge amounts of data and the need for transforming such data into knowledge influences towards the attraction of IT industry in data mining. During the early years of the development of computer techniques for business, IT professionals were concerned with designing databases to store the data so that information could be easily and quickly accessed. The restrictions are storage space and the speed of retrieval of the data. Needless to say, the activity was restricted to a very few, highly qualified professionals. Then came an era when Database Management System simplified the task. Thus almost any business such as small, medium or large scale began using computers for day - to- day activities. Now what is the use of all this data? Up to the early 1990's the answer to this was "NOT much". No one was really interested in utilizing data, which was accumulated during the process of daily activities. As a result a new discipline in Computer Science, Data Mining gradually evolved. Data mining is becoming a pervasive technology in activities as diverse as using historical data to predict the success of a marketing campaign, looking for patterns in financial transactions to discover illegal activities or analyzing genome sequences. This paper deals with the application of data mining in various fields in our day to day life.**

Keywords-Data Mining, Targeted Marketing, Market Based Analysis, Customer Relations

I. INTRODUCTION

Data Mining – An Overview

Data mining refers to extracting knowledge from large amounts of data. The data may be spatial data, multimedia data, time series data, text data and web data. Since Data mining is a young discipline with wide and diverse applications. In this paper we will discuss a few application domains of data mining such as Science and Engineering, Banking, Business, Telecommunication and Surveillance.

Data mining is the process of extraction of interesting, nontrivial, implicit, previously unknown and potentially useful patterns or knowledge from huge amounts of data. It is the set of activities used to find new, hidden or unexpected patterns in data or unusual patterns in data. Using information contained within data warehouse, data mining can often provide answers to questions about an organization that a decision maker has previously not thought to ask.

About-[1] Department of Computer Science, Government Arts College (Women) , Krishnagiri- 635 001, India
(e-mail;sundaripalanisamy@yahoo.co.in)
About-[2] Department of Computer Science, Government Arts College. (Men) , Krishnagiri- 635 001, India

- Which products should be promoted to a particular customer? – Targeted Marketing
- What is the probability that a certain customer will leave for a competitor? – Customer Relationship Management
- What is the appropriate medical diagnosis for this patient? – Bio medical
- What is the likelihood that a certain customer will default or pay back a loan? – Banking
- Which products are bought most often together? – Market Basket Analysis
- How to identify fraudulent users in telecommunication industry? – Fraudulent pattern analysis

These types of questions can be answered quickly and easily if the information hidden among the huge amount of data in the databases can be located and utilized. We will discuss about the applications of data mining in the following paragraphs.

II. APPLICATIONS OF DATA MINING

Although a large variety of data mining scenarios can be discussed, for the purpose of this paper the applications of data mining are divided into the following categories:

- Science and Engineering
- Business
- Banking
- Telecommunication
- Spatial data mining
- Surveillance

II. (A) Science and Engineering

The data mining has been widely used in area of science and engineering, such as bioinformatics, genetics, medicine, education and electrical power engineering.

i) Biomedical and DNA Data analysis

The past decade has seen an explosive growth in biomedical research, ranging from the development of new pharmaceuticals and in cancer therapies to the identification and study of human genome by discovering large scale sequencing patterns and gene functions. Recent research in DNA analysis has led to the discovery of genetic causes for many diseases and disabilities as well as approaches for disease diagnosis, prevention and treatment. It is challenging to identify particular gene sequence patterns that play roles in various diseases. DNA data analysis is done in the following ways.[5]

- Semantic integration of heterogeneous, distributed genome databases
- Similarity search and comparison among DNA sequences
- Identification of co occurring gene sequences
- Path analysis includes linking genes to different stages of disease development
- Visualization tools and genetic data analysis
- The data mining technique that is used to perform this task is known as Multifactor Dimensionality Reduction.[3]

In adverse drug reaction surveillance, the Uppsala Monitoring Centre has, since 1998, used data mining methods to routinely screen for reporting patterns indicative of emerging drug safety issues in the WHO global database of 4.6 million suspected adverse drug reaction incidents.[7] Recently, similar methodology has been developed to mine large collections of electronic health records for temporal patterns associating drug prescriptions to medical diagnoses.[8]

ii) Education

The other area of application for data mining in science/engineering is within educational research, where data mining has been used to study the factors leading students to choose to engage in behaviors which reduce their learning and to understand the factors influencing university student retention.[6] A similar example of the social application of data mining is its use in expertise finding systems, whereby descriptors of human expertise are extracted, normalized and classified so as to facilitate the finding of experts, particularly in scientific and technical fields. In this way, data mining can facilitate Institutional memory.

iii) Electrical power engineering

In the area of electrical power engineering, data mining techniques have been widely used for condition monitoring of high voltage electrical equipment. The purpose of condition monitoring is to obtain valuable information on the insulation's health status of the equipment. Data clustering such as Self-Organizing Map (SOM) has been applied on the vibration monitoring and analysis of transformer On-Load Tap-Changers(OLTCS). Using vibration monitoring, it can be observed that each tap change operation generates a signal that contains information about the condition of the tap changer contacts and the drive mechanisms. Obviously, different tap positions will generate different signals. However, there was considerable variability amongst normal condition signals for the exact same tap position. SOM has been applied to detect abnormal conditions and to estimate the nature of the abnormalities.[4]

Data mining techniques have also been applied for Dissolved Gas Analysis (DGA) on power transformers. DGA, as a diagnostics for power transformer, has been available for many years. Data mining techniques such as SOM has been applied to analyze data and to determine trends which are not obvious to the standard DGA ratio techniques such as Duval Triangle.[4]

Data mining technique is used to an integrated-circuit production line[2]. The data mining technique is applied in decision analysis to the problem of die-level functional test. Experiments demonstrate the ability of applying a system of mining historical die-test data to create a probabilistic model of patterns of die failure which are then utilized to decide in real time which die to test next and when to stop testing. This system has been shown, based on experiments with historical test data, to have the potential to improve profits on mature IC products.

b) Banking

Banking data mining applications may, for example, need to track client spending habits in order to detect unusual transactions that might be fraudulent. Most banks and financial institutions offer a wide variety of banking services (such as checking, saving, and business and individual customer transactions), credit (such as business, mortgage, and automobile loans), and investment services (such as mutual funds) [5]. It has also offer insurance services and stock services. For example it can also help in fraud detection by detecting a group of people who stage accidents to collect on insurance money. The following methods are used for financial data analysis.

- Loan payment prediction and customer credit policy analysis
- Classification and clustering of customers for targeted marketing
- Detection of money laundering and other financial crimes

c) Business

Retail industry collects huge amount of data on sales, customer shopping history, goods transportation and consumption and service records and so on. The quantity of data collected continues to expand rapidly, especially due to the increasing ease, availability and popularity of the business conducted on web, or e-commerce. Retail industry provides a rich source for data mining. Retail data mining can help identify customer behavior, discover customer shopping patterns and trends, improve the quality of customer service, achieve better customer retention and satisfaction, enhance goods consumption ratios design more effective goods transportation and distribution policies and reduce the cost of business [5]. A few examples of data mining in the retail industry are as follows.

- Design and construction of data warehouses based on benefits of data mining
- Multidimensional analysis of sales, customers, products, time and region:

The multi feature data cube is a useful data structure in retail data analysis.

Another example of data mining, often called the market basket analysis, relates to its use in retail sales. If a clothing

store records the purchases of customers, a data-mining system could identify those customers who favors silk shirts over cotton ones. Although some explanations of relationships may be difficult, taking advantage of it is easier. The example deals with association rules within transaction-based data. Not all data are transaction based and logical or inexact rules may also be present within a database. In a manufacturing application, an inexact rule may state that 73% of products which have a specific defect or problem will develop a secondary problem within the next six months.

Market basket analysis has also been used to identify the purchase patterns of the Alpha consumer. Alpha Consumers are people that play key roles in connecting with the concept behind a product, then adopting that product, and finally validating it for the rest of society. Analyzing the data collected on these type of users has allowed companies to predict future buying trends and forecast supply demands.

Data Mining is a highly effective tool in the catalog marketing industry. Catalogers have a rich history of customer transactions on millions of customers dating back several years. Data mining tools can identify patterns among customers and help identify the most likely customers to respond to upcoming mailing campaigns.

- Analysis of the effectiveness of sales campaigns:
- Customer retention – analysis of customer loyalty

There are a wide variety of data mining applications available, particularly for business uses, such as Customer Relationship Management (CRM). Goods purchased at different periods by the same customers can be grouped into sequences. Sequential pattern mining can be used to investigate changes in customer consumption and suggest adjustments on the pricing and variety of goods in order to help retain customers and attract new customers. These applications enable marketing managers to understand the behaviors of their customers and also to predict the potential behavior of prospective customers. A data mining technique may assist the prediction of future customer retention. For example, a company may decide to increase prices, and could use data mining to predict how many customers might be lost for a particular percentage increase in product price.

Data mining can also be helpful to human-resources departments in identifying the characteristics of their most successful employees. Information obtained, such as universities attended by highly successful employees, can help HR focus recruiting efforts accordingly. Additionally, Strategic Enterprise Management applications help a company translate corporate-level goals, such as profit and margin share targets, into operational decisions, such as production plans and workforce levels.[1]

d) Telecommunication

The telecommunication industry offers local and long distance telephone services to provide many other comprehensive communication services including voice, fax, pager, cellular phone, images, e-mail, computer and web data transmission and other data traffic. The integration of telecommunication, computer network, Internet and numerous other means of communication and computing are underway. Moreover, with the deregulation of the telecommunication industry in many countries and the development of new computer and communication technologies, the telecommunication market is rapidly expanding and highly competitive. This creates a great demand from data mining in order to help understand business involved, identify telecommunication patterns, catch fraudulent activities, make better use of resources, and improve the quality of service.

e) Spatial data mining

Spatial data mining is the application of data mining techniques to spatial data. It follows along the same functions in data mining, with the end objective to find patterns in geography. So far, data mining and Geographic Information Systems (GIS) have existed as two separate technologies, each with its own methods, traditions and approaches to visualization and data analysis. Particularly, most contemporary GIS have only very basic spatial analysis functionality. The immense explosion in geographically referenced data occasioned by developments in IT, digital mapping, remote sensing, and the global diffusion of GIS emphasizes the importance of developing data driven inductive approaches to geographical analysis and modeling.

Data mining, which is the partially automated search for hidden patterns in large databases, offers great potential benefits for applied GIS-based decision-making. Recently, the task of integrating these two technologies has become critical, especially as various public and private sector organizations possessing huge databases with thematic and geographically referenced data begin to realize the huge potential of the information hidden there. Among those organizations are:

Offices requiring analysis or dissemination of geo-referenced statistical data.

Public health services searching for explanations of disease clusters.

Environmental agencies assessing the impact of changing land-use patterns on climate change.

Geo-Marketing companies doing customer segmentation based on spatial location.

f) Surveillance

Data Mining is used by intelligence agencies like FBI and CIA to identify threats of terrorism. After the 9/11 incident it has become one of the prime means to uncover terrorist plots. However this led to concerns among the people as data collected for such works undermines the privacy of a large number of people.

Two plausible data mining techniques in the context of combating terrorism include "pattern mining" and "subject-based data mining".

i) Pattern mining

"Pattern mining" is a data mining technique that involves finding existing patterns in data. Pattern mining is a tool to

identify terrorist activity, the National Research Council provides the following definition: "Pattern-based data mining looks for patterns (including anomalous data patterns) that might be associated with terrorist activity — these patterns might be regarded as small signals in a large ocean of noise."[9][10][11] Pattern Mining includes new areas such a Music Information Retrieval (MIR) where patterns seen both in the temporal and non temporal domains are imported to classical knowledge discovery search techniques.

ii) Subject-based data mining

"Subject-based data mining" is a data mining technique involving the search for associations between individuals in data. In the context of combating terrorism, the National Research Council provides the following definition: "Subject-based data mining uses an initiating individual or other datum that is considered, based on other information, to be of high interest, and the goal is to determine what other persons or financial transactions or movements, etc., are related to that initiating datum."[9]

g) Text Mining and Web Mining

Text mining is the process of searching large volumes of documents from certain keywords or key phrases. By searching literally thousands of documents various relationships between the documents can be established.

An extension of text mining is web mining. Web mining is an exciting new field that integrates data and text mining within a website. Web serves as a huge, widely distributed, global information service center for news, advertisements, consumer information, financial management, education, government, e- commerce and many other information services. It enhances the web site with intelligent behavior, such as suggesting related links or recommending new products to the consumer. Web mining is especially exciting because it enables tasks that were previously difficult to implement. They can be configured to monitor and gather data from a wide variety of locations and can analyze the data across one or multiple sites. For example the search engines work on the principle of data mining.

III. Need of Data Mining

The massive growth of data is due to the wide availability of data in automated form from various sources as WWW, Business, science, Society and many more. Data is useless, if it cannot deliver knowledge. That is why data mining is gaining wide acceptance in today's world. A lot has been done in this field and lot more need to be done.

IV. Conclusion

Since data mining is a young discipline with wide and diverse applications, there is still a nontrivial gap between general principles of data mining and domain specific, effective data mining tools for particular applications. The aim of the paper is the study of application domains of Data Mining such as science and engineering, banking, business and telecommunication. Although data mining is a young field with many issues that still need to be researched in depth. The diversity of data, data mining tasks and approaches poses many challenging research issues in data mining. The design of data mining languages, the development of efficient and effective data mining methods, the construction of interactive and integrated data mining environments and the application of data mining techniques to solve large application problems are important tasks for data mining researchers.

V. References

1) Ellen Monk, Bret Wagner (2006). Concepts in Enterprise Resource Planning, Second Edition. Thomson Course Technology, Boston, MA. ISBN 0-619-21663-8. OCLC 224465825.
2) Tony Fountain, Thomas Dietterich & Bill Sudyka (2000) Mining IC Test Data to Optimize VLSI Testing, in Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. (pp. 18-25). ACM Press.
3) Xingquan Zhu, Ian Davidson (2007). Knowledge Discovery and Data Mining: Challenges and Realities. Hershey, New Your. pp. 18. ISBN 978-159904252-7.
4) a b A.J. McGrail, E. Gulski et al.. "Data Mining Techniques to Asses the Condition of High Voltage Electrical Plant". CIGRE WG 15.11 of Study Committee 15.
5) Jiawei Han & Micheline Kamber. (2001) Data Mining: Concepts and Techniques , Morgan Kaufmann publishers, CA,USA.
6) J.F. Superby, J-P. Vandamme, N. Meskens. "Determination of factors influencing the achievement of the first-year university students using data mining methods". Workshop on Educational Data Mining 2006.
7) Bate A, Lindquist M, Edwards IR, Olsson S, Orre R, Lansner A, De Freitas RM. A Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol. 1998 Jun;54(4):315-21.
8) Norén GN, Bate A, Hopstadius J, Star K, Edwards IR. Temporal Pattern Discovery for Trends and Transient Effects: Its Application to Patient Records. Proceedings of the Fourteenth International Conference on Knowledge Discovery and Data Mining SIGKDD 2008, pages 963-971. Las Vegas NV, 2008.

9) a b National Research Council, Protecting Individual Privacy in the Struggle Against Terrorists: A Framework for Program Assessment, Washington, DC: National Academies Press, 2008.

10) R. Agrawal et al., Fast discovery of association rules, in Advances in knowledge discovery and data mining pp. 307-328, MIT Press, 1996.

11) Stephen Haag et al. (2006). Management Information Systems for the information age. Toronto: McGraw-Hill Ryerson. pp. 28. ISBN 0-07-095569-7. OCLC 63194770.