Data Mining in Research Keypad of Discoverer

Biplab Biswas
Dept. of Computer Science
Sripat Singh College, Jiaganj, Murshidabad, India
Email: biplabbiswas@gmail.com

Abstract – Knowledge plays a key role in human lives. Data Mining is an analytic process designed to explore data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prophecy - and prophetic data mining is the most common type of data. The process of data mining consists of three stages: (1) the initial exploration, (2) model building or pattern identification with validation/verification, and (3) deployment. In short, data mining is the process of deriving patterns from large databases. Due to the importance of extracting knowledge/information from the large data storehouse, data mining has become an essential component in various fields of human life. Advancements in Statistics, Machine Learning, Artificial Intelligence, Pattern Recognition and Computation capabilities have evolved the present day’s data mining applications and these applications have enriched the various fields of human life. This data is used for machine learning and predictive analysis. Current sphere of data mining ranges from theoretical work on the principles of learning and mathematical representations of data to building advanced engineering systems that perform information filtering on the web, find genes in DNA sequences, help understand trends and anomalies in economics and education and detect network intrusion. Data mining is also a promising computational paradigm that enhances traditional approaches to discovery and increases the opportunities for breakthroughs in the understanding of complex physical and biological systems. The ever increasing complexities in various fields and improvements in technology have posed new challenges to data mining; the various challenges include different data formats, data from disparate locations, advances in computation and networking resources, research and scientific fields, ever growing business challenges etc. Advancements in data mining with various integrations and implications of methods and techniques have shaped the present data mining applications to handle the various challenges of the modern world.

Keywords – Data Mining, Data repositories, Artificial Intelligence, Pattern Recognition.

I. INTRODUCTION

Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. It is the process of analyzing data from different perspectives and summarizing it into useful information - that can be used to increase revenue, cut cost, or both. We are living in an age of information - because we believe that information leads to power and success, and thanks to stylish or advance technologies such as computers, Mobiles, satellites, etc., from where we have been collecting incredible amounts of information.

After advent of computers and resources for mass digital storage, we started collecting and storing all kinds of data, counting on the power of computers to help sort through various combination of information. Unfortunately, the collections of those massive amalgamate data stored on different structures very rapidly became overpowering. This initial chaos gives birth of structured databases and database management systems (DBMS). In this age, we face more informations than we can handle:- from business transactions and scientific data, from text reports and military intelligence. Dealing with huge collections of data, we have now created new types of database to help us make better decision-making choices. These needs are automatic summarization of data, extraction of the “real meaning” of information stored, and the innovation of patterns in raw data.

It is concerned with finding hidden relationships present in business data to allow business to make predictions for future use. It is the process of data-driven extraction of not so obvious, but useful information from large databases. Data mining has emerged as a key business intelligence technology.

DM is the set of activities used to find new, hidden, or unexpected patterns in data. It extraction of hidden predictive information from large database, is powerful new technology with great potential to help companies focus on the most important information in their data warehouses. The automated, prospective analyses offered by data mining move beyond the analyses of past event provided by demonstration tools typical of decision support systems. It is the process of data-driven extraction of not so obvious, but useful information from large database. Data mining has emerged as a key business intelligent technology. But ultimately some question arising that- where can it be useful? And how does it work?

II. PURPOSE OF DATA MINING

Let us discuss the purpose of data mining with point of sale system (POS). Usually supermarkets make use of a POS system that collects data from each item that is purchased. The POS system collects data on the item brand name, category, size, time and date of the purchase and at what price the item purchased. In addition, the supermarket usually has a customer rewards program.
Which is also an input to the POS system. This information can directly link the products purchased with an individual. All this data for every purchase made for years and years is stored in a database in a computer by the supermarket. But here they have database with millions of records. For that some questions are arrive like:

- What will they do with huge data?
- How do they use this data to forecast or control their business activities? etc.

All questions have only one solution is Data mining. Using data mining techniques or algorithm, we can uncover trends, statistical correlations, relationships and patterns that can help our business become more efficient, effective and streamlined.

According to that, the supermarket can now figure out brand as sell the most, what time of the day, week, month or year is most busiest, what products do consumers buy along with certain items. For instance, if a person buys white bread, what other item would they be inclined to buy? Typically we can find its peanut butter and jelly. There is so much good information that a supermarket can use just by data mining their own data that they have collected. The most recent application of DM is in the unstable growth area of customer relationship management. DM and customer relationship management software enable users to analyze large databases to solve business-decision problems. Like statistics, data mining is not a business solution; it is a technology. CRM, on other hand, involves turning information in a database into a business decision that drives interactions with customers.

### III. DATA MINING IN E-COMMERCE

Even more recent application of data mining techniques for collected data from visitors to an E-commerce website. Companies venturing into E-commerce have a dream. By analyzing the tracks people made through their web site, they can better optimize the design of their sites to maximize sales. Information about customers and their purchasing habits enables companies to initiate E-Mail campaigns and other activities that result in sales. Good modules of customers preferences, need, desires and behaviors help companies stimulates the personal relationship that businesses and their clientele had in the good old days.

By analyzing customer paths through the data, vendors hope to personalize their interactions with customers and prospects. Companies will customize the home page—each customer sees, the responses to customer requests, and the recommendations of items to purchase. Vendors can also generate a list of related products.

The business benefits’ of this customer intelligence are potentially enormous. For those sites providing such personalization, the number of people who come to a site and purchase will increase and the average amount per purchases will rise, resulting in a dramatic increase in profitability or at least that is what the dream says.

The reality is that achieving this goal of personalization the increase profits is difficult and expensive; however, it is possible. First, to be of any use at all, click-stream data, the data collected by monitoring the various "click" patterns of a typical user when navigating through the web require enormous amount of labor-intensive preprocessing. Even then extracting meaning is still difficult. Second many customers are reluctant to have vendor’s tracks what they do online. Their concern is so great that the Government is actively considering privacy regulation to limit web tracking. Nonetheless, DM is emerging as a major force in deep data analysis.

### IV. DM'S GROWTH IN POPULARITY

Several reasons can be mentioned in support of the growing popularity of DM. Maybe the single greatest reason for popularity is the ever-increasing volume of data that require processing: the amount of data collects each day by business and organizations according to their function. A report from the GTE Research Center (2000) suggests that scientific and academic organizations store approximately 1 terabyte of new data each day, even though the academic community is not the leading supplier of new data worldwide.

Another reason for the growing popularity of DM is an increasing awareness of the inadequacy of the human brain to process data, particularly in situations involving multi-factorial dependencies or correlations.

Finally a third reason for the grouping popularity of DM is the increasing affordability of machine learning. An automated DM system can operate at a much lower cost than professional statisticians. Although DM has not entirely eliminated human involvement, alternately it allows humans to better manage the process.

### V. HOW DOES DATA MINING WORK

Data mining is still in its infancy, companies in a wide range of industries—include finance, health care, manufacturing, transportation—are already using data mining tools and techniques to take advantages of historical data.

Logic of data mining is based on modeling. A model based on data from situations where the answer is known and then applying the model to other situations where the answers are not known. The computational power to automate modeling techniques to work directly on the data.

Let’s see an example of building a model, consider the marketing director for a telecommunications company, they would like to focus his marketing and sales efforts on segments of the population most likely to become big users of long-distance service. They know a lot about their customers, but it is impossible to detect the common characteristics of their customers. From this existing database of customers, which contains information such as...
age, sex, credit history, income, pin code, occupation, etc. they can use data mining tools, such as neural networks, to identify the characteristics of those customers who make lots of long-distance calls. For instance, he might learn that his best customers are unmarried females between the ages of 21 to 40 who earn in excess of Rs-60000 per year etc. by this according their model they find out high-value customers, and they would budget his marketing effort accordingly.

According that we can say, data mininhg is the task of discovering interesting patterns from large amounts of data where the data can be stored in databases, data warehouses or other information warehouses.

VI. DATA MINING TECHNOLOGIES

The analytical techniques used in data mining are often well-known mathematical algorithms and techniques. Some new applications and techniques to general business problems made possible to increased availability of data, expensive storage and processing power. Also, the use of graphical interface has led to tools becoming available that business expert can easily use.

Some of the techniques are given below:-

a) Artificial neural networks – Nonlinear predictive models that learn through training and resemble biological neural network in structure.

b) Decision tree – tree-shaped structures that represent sets of decisions. These decisions generate rules of the classification of a dataset.

c) Rule induction – the extraction of useful if-then rules from databases on Statistical significance.

d) Genetic algorithms: Optimization techniques based on the concepts of genetic combination, mutation, and natural selection.

e) Nearest neighbour – A classification technique that classifies each record based on the records most similar to it in a historical database.

Some data mining applications usually used by industries are-

- Web site design
- Diagnosing medical problems
- Inventory management
- Relationship management
- Conduct on-line marketing
- Telecomunications
- Banking
- Biometrics
- Insurance … etc

VII. DATA WAREHOUSE

The construction of data warehouse, which involves data cleaning and data integration, can be viewed as an important preprocessing step for data mining. Data warehouse have been define in many ways, making it difficult to formulate a rigorous definition. A data warehouse refers to a database that is maintained separately from an organization’s operational databases. Data warehouse systems allow for the integration of a variety of application systems.

According to W.H.Inmon, a leading architecture in the construction of data warehouse systems, “a data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision making process”.

VIII. DESIGN OF A DATA WAREHOUSE

To design an effective data warehouse one need to understand and analyze business needs and construct a business analysis framework. The construction of a large and complex information system can be viewed as the construction of a large and complex building, for which the owner, architect, and builder have different views. According that a data warehouse must be consider of four different views-

a) Top-down views: it allows the selection of the relevant information necessary for the data warehouse.

b) Data source view: it exposes the information stored, and manage by operational. Data sources are often modeled by traditional data modeling techniques, like E-R model(Entity- Relational model), or CASE (Computer-Aided Software Engineering) tools.

c) Data warehouse view: it includes fact tables and dimension tables. It presents the information that is stored inside the data warehouse, including precalculated totals and counts.

d) Business query view: this view is the perspective of data in the data warehouse from the viewpoint of the end user.

According the point of views of software engineer’s- the design and constructed of a data warehouse may consist of the following steps- planning, requirement analysis, feasibility study, ware house design, coading and testing and deployment of data warehouse. The waterfall method and spiral method are the two methodologies by which software can design. The waterfall method and spiral method performs a structured and systematic analysis at each step before proceeding to the next. The spiral method involve the rapid generation of increasingly functional systems, with short intervals between successive releases.

IX. DESIGN OF A DATA WAREHOUSE

Data warehouses often adopt three-level architecture-
In above discussion, the introduction to the data warehousing techniques is outlined. It includes the basic concept of data warehousing, its architecture and components. The three level architecture also demonstrate the need for data mining. How data mining works in various fields and the analytical data mining techniques also mention.

**REFERENCES**


Fig. 1. Three-level architecture of Data Warehouse

1) The bottom level is a warehouse database server that is a relational database system. Data from operational databases and external sources are extracted by program interfaces known as gateways. It is supported by the DBMS and allows client program to generate SQL code. Like, ODBC (Open Database Connection), OLE-DB (Open Linking and Embedding for Databases), and JDBC (Java Database Connection).

2) The middle level is an OLAP that is typically implemented using either relational OLAP (ROLAP) or multidimensional OLAP (MOLAP).

3) Top level is a client, it contains query and reporting tools, analysis tools, data mining tools.

According to the architecture, warehouse collects all of the information about subjects spanning the entire organization. It provides wide data integration from operational systems or external information. It contains detailed data that range from few gigabytes to hundreds of gigabytes or beyond (Tera Byte, Peta Byte, Exa Byte, Zetta Byte, Yotta Byte).

Data warehouse is non-volatile types means read-only environment. Data can not be updated after loaded; it is update when it comes into the operational environment. So, no need to worry about uploaded data from abnormal update. Data update in this environment by refreshing or periodic loading of new data from working environment; for example, online placement system (upload Bio-Data from employee and update regularly through refreshing or load new data), cloud computing, social network (facebook).