IMPLICATIONS OF DATA MINING IN DIGITAL LIBRARY ENVIRONMENT

R N Mishra

Abstract

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help the Libraries and Information Centers to focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing Libraries and Information Centers managers to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can resolve the problems in Libraries and Information Centers especially in a digital environment. Data Mining has become imminent in view of the abundant growth of digital resources for its management, organization, and retrieval and dissemination of the same to the right users instantly to satisfy the varied need of the users due to inter-disciplinary research. Data Mining, Viability of Data Mining, Process involved in Data Mining, etc. have been discussed in this paper.

Keywords : Data Mining, KDD, Artificial Neural Networks, Sequential Pattern, Modeling building

1. Introduction

Data Mining has been defined in multifarious ways with different implications by various organizations and stalwarts. Data Mining can be referred to a heave of accumulated data in various areas. It can be equated with the term such as knowledge discovery and this infect is a process which is associated with analyzing data from different perspectives and abridging it into useful information. The information can be used to increase revenue, cuts costs, or both. Software has been specifically designed known as Data mining software which has a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases (www.anderson.ucla.edu). Data Mining is a terminology which refers practically to Data Extraction from a heave of data available in electronic form.

According to Wikipedia, Data Mining is a nontrivial extraction of implicit and potentially useful information from data and the science of extracting useful information from large data sets or databases (http://en.wikipedia.org/wiki/Data_mining). Data mining involves sorting through large amounts of data and picking out relevant information. It is usually used by Business Intelligence Organizations, and financial analysts, but is increasingly used in the sciences to extract information

from the enormous data sets generated by modern experimental and observational methods. Library and Information Service being one of the outstanding service field can well be accommodated with the arena of Data Mining as a good quantum of data are prevalent in the Libraries and Information Centers especially in an electronic environment.

Data Mining initially emerged from the business arena and has recently developed its importance to governmental agencies due to security issues in the US (US Government Accountability Office, 2005. The emphasis on Gabriel Gomez consumers indicates how libraries may well come to deploy data mining in the future, that is to more accurately predict the needs of library users as if they were consumers. Data mining might well be used in library settings for the creation of profiles to spot user patterns, and ultimately, to predict user behavior (Gabriel Gomez; 2007; 519-28). Data mining identifies trends within data that go beyond simple analysis. Through the use of sophisticated algorithms, users have the ability to identify key attributes of business processes and target opportunities.

The term Data Mining is often used to apply to the two separate processes of knowledge discovery and prediction. Knowledge discovery provides explicit information that has a readable form and can be understood by a user. Forecasting, or predictive modeling provides predictions of future events and may be transparent and readable in some approaches (e.g. rule based systems) and opaque in others such as neural networks. Moreover, some Data Mining systems such as neural networks are inherently geared towards prediction and pattern recognition, rather than knowledge discovery.

Towards the end of 1980s machine learning methods for searching were started as a means of beyond the fields of computing and artificial intelligence, which were employed in database marketing applications where the available databases were used for elaborate and specific marketing campaigns. The term Knowledge Discovery in Databases (KDD) was first time coined to describe all those methods which aimed to find relations and regularity among the observed data (Giudici; 2005; p.2). Gradually due to technological advances in data capture, processing power along with data transmission and storage capabilities, the large organizations like libraries and information centers in electronic environment especially have started integrating their various databases in to data warehouses and this can well be defined as a process of Centralized Data Management including its retrieval. It may be mentioned that, the term Data Warehouses is relatively new term which can be applied to service organizations like libraries and information centers. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. Centralization of data is needed to maximize user access and analysis. Dramatic technological advances are making this vision a reality for many companies. And, equally dramatic advances in data analysis software are allowing users to access this data freely. The data analysis software is what supports data mining.

According to Marketing Dictionary Data Mining concerns to Extraction of customer information from a database by utilizing software that can isolate and identify previously unknown patterns or trends in large amounts of data. There are a variety of data mining techniques that reveal different types of patterns. According to Intelligence Encyclopedia Data Mining Data mining refers to the statistical analysis techniques used to search through large amounts of data to discover trends or patterns (www.answers.com/topic/data-mining).

2. Viability of Data Mining

Data Retrieval, like Data Mining, extracts interesting data and information from archives and various databases. The difference between Data Retrieval and Data Mining is that, unlike Data Mining the criteria for extracting information are decided beforehand so that they are exogenous from the extraction its-self. Further, Data Mining is different from Data Retrieval as it searches the relationships and associations between the unknown phenomena.

To mention a brief history, Wal-Mart named after Sam Walton of USA happens to be the one of the biggest American Public Corporation was the first organization to use the Data Mining technology for its grocery and consumable business and this technique are primarily associated with many national and international agencies with a strong consumer focus such as, retail, financial, communication and marketing. This also enables the agencies to determine the relationships among internal factors such as products, service, staff and the external factors such as competition etc.

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. The neural networks based on the concept of Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems (http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html). Generally four types of relationships can be sought in the Data Mining which can be applied to the Libraries especially in an electronic environment.

- **Classes :** Stored or accumulated pool of data is used to locate data for predetermined groups. The Library requires grouping the data according to the type of clienteles such as, engineers, doctors, and lawyers etc. who require data for specific purpose. The data can be stored in electronic form for more feasibility and accessibility. This information could be used to increase traffic by having daily specials.

- **Clusters:** Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.
- **Associations :** Data can be mined to identify associations. The beer-diaper example is an example of associative mining.
- **Sequential patterns :** Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

3. Data Mining Process

A series of activities or processes are involved in Data Mining and this is an analytic process which is designed to explore data from a good chunk of data and this is involved in searching process 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 prediction and this predictive data mining is the most common type of data mining which has the direct business applications. It will not be out of place to mention that, the Predictive Data Mining is invariably employed to identify data mining projects with an ambition to recognize a statistical or neural network model or set of models that can be used to predict some response of interest. The process of data mining consists of seven phases such as (Giudici; 2005; p.6),

- Objective of the analysis;
- Selection, organization of the data;
- Exploratory analysis and transformation of data;
- Specification of the statistical methods to be employed while analyzing the data;
- Analysis of the data on chosen method;
- Evaluation and comparison of the methods and choice for final method; and
- Interpretation for decision making process.

In brief, Data Mining involves three stages for exploration such as,

- Initial exploration;
- Model building or pattern identification with validation/verification; and
- Deployment (i.e., the application of the model to new data in order to generate predictions).

The first stage of Exploration normally is commensurate with data preparation which involves cleaning of data by removing the redundant elements which follows data transformations. Thereafter, subsets of records are being selected where each case of data sets figures with large numbers of variables ("fields"). Depending upon the statistical methods employed, operation part is carried out to bring the number of variables to a manageable range

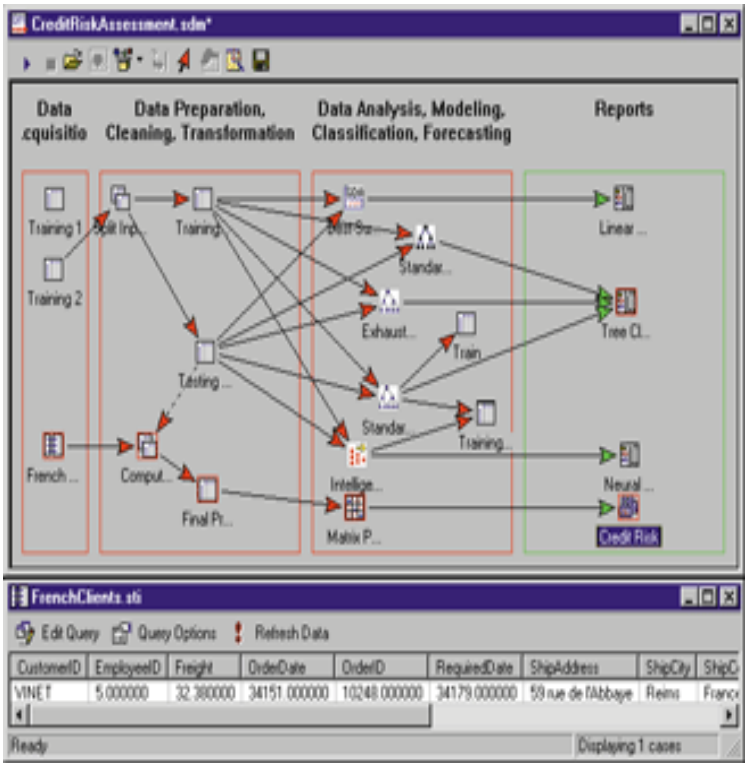

Fig- 1- Data Mining Process

Then, depending on the nature of the analytic problem, this first stage of the process of data mining may involve anywhere between a simple choice of straightforward predictors for a regression model, to elaborate exploratory analyses using a wide variety of graphical and statistical methods in order to identify the most relevant variables and determine the complexity and/or the general nature of models that can be taken into account in the next stage.

Model building and validation is the second stage in Data Exploration which involves considering various models and choosing the best one based on their predictive performance (i.e., explaining the variability in question and producing stable results across samples). This may sound like a simple operation, but in fact, it sometimes involves a very elaborate process. There are a variety of techniques developed to achieve that goal - many of which are based on so-called "competitive evaluation of models," that is, applying different models to the same data set and then comparing their performance to choose the best. These techniques - which are often considered the core of predictive data mining - include: Bagging (Voting, Averaging), Boosting, Stacking (Stacked Generalizations), and Meta-Learning.

The third and final stage relates to Deployment which involves using the selected model best suit to the previous stage and applying it thereby, to new data in order to generate predictions or estimates of the expected outcome.

4. Elements of Data Mining

The elements of Data Mining can be broadly grouped under six headings such as,

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

5. Levels of Analysis

Levels of analysis involved in Data Mining can be summarized as

- Artificial Neural Networks;
- Genetic Algorithms;
- Decision Trees;
- Nearest Neighbor Method;
- Rule Induction; and
- Data Visualization.

5.1 Artificial Neural Networks

It is a Non-linear predictive model that learns through training and resemble. It is associated with a biological neural network in structure. An Artificial Neural Network (ANN), often just called a "Neural Network" (NN), is a mathematical model or computational model based on biological neural networks. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase (http://en.wikipedia.org/wiki/Artificial_neural_network#Models).

5.2 Genetic Algorithms

It is a process of optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

5.3 Decision Trees

Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID), CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments are a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

5.4 Nearest Neighbor Method

It is involved with a technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k 1). Sometimes it is called the k-nearest neighbor technique.

5.5 Rule Induction

The extraction of useful if-then rules from data based on statistical significance.

5.6 Data Visualization

The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

6. Use of Data Mining in Libraries

In the digital scenario, the Libraries and Information Centers use to accumulate more and more digital data, process, manage and archive every day. Due to proliferation of information wealth in the service segments such as Libraries Data Mining has become imminent for effective and instant retrieval of information. Data mining which has got a lot of implications as a practice in the Libraries and Information Centers can be operated with multifaceted understandings regarding its use. The ALA website has numerous things written on data mining. For example: Data mining is the practice of aggregating information about consumers' preferences and interests from a variety of sources, including cookies, stealth data software, voluntary purchases, and mailing lists, with the purpose of creating comprehensive profiles. Most often the profiles are used for targeted advertisements. But federal and local governments are also increasingly relying on data mining to assemble profiles to investigate various criminal and fraudulent activities (Tools Used for Collecting Online Data, 2006, para. 5). Such complex analysis is dependent upon a key distinction. Data mining is not easily done from systems that currently manage library services.

However, Kevin Cullen notes that Data mining is performed in a data warehouse. While an operational database system like an integrated library system (ILS) is optimized for processing transactions (circulation, purchases, cataloging, etc), a data warehouse is optimized for analysis. This makes it easier to find patterns and avoid bogging down the transactional system (Cullen, 2005, p. 31).

The development of computational algorithms is other dimensions of Data Mining which is used for the identification or extraction of structure from data. This has become essential to understand, analyze the data. Tasks supported by data mining include prediction, segmentation, dependency modeling, summarization, and change and deviation detection. Database systems have brought digital data capture and storage to the mainstream of data processing, leading to the creation of large data warehouses. These are databases whose primary purpose is to gain access to data for analysis and decision support. Traditional manual data analysis and exploration requires highly trained data analysts and is ineffective for high dimensionality (large numbers of variables) and massive data sets.

A data set can be viewed abstractly as a set of records, each consisting of values for a set of dimensions (variables). While data records may exist physically in a database system in a schema that spans many tables, the logical view is of concern here. Databases with many dimensions pose fundamental problems that transcend query execution and optimization. A fundamental problem is query formulation: How is it possible to provide data access when a user cannot specify the target set exactly, as is required by a conventional database query language such as SQL (Structured Query Language)? Decision support queries are difficult to state. For example, which records are likely to represent fraud in credit card, banking, or telecommunications transactions? Which records are most similar to records in table A but dissimilar to those in table B? How many clusters (segments) are in a database and how are they characterized? Data mining techniques allow for computer-driven exploration of the data, hence admitting a more abstract model of interaction than SQL permits (http://www.answers.com/topic/data-mining).

Data mining techniques are fundamentally data reduction and visualization techniques. As the number of dimensions grows, the number of possible combinations of choices for dimensionality reduction explodes. For an analyst exploring models, it is not practically possible to go through the various ways of projecting the dimensions or selecting the right sub samples (reduction along columns and rows). Data mining is based on machine-based exploration of many of the possibilities before a selected reduced set is presented to the analyst for feedback.

Exploring and analyzing data mining implies digging through heave of data to unfurl patterns and relationships contained within the organization including business activity. Data mining can be done manually by slicing the data until a right pattern becomes obvious. It can also be done with

programs that can analyze the data automatically. Data mining has become an important part for maintaining the Customer Relationship Management (CRM).

Data mining is an especially powerful tool in the examination and analysis of huge databases. With the advent of the Internet, vast amounts of data are accumulating. As well, the amount of data that can be generated from a single scientific experiment where stretches of DNA are affixed to a glass chip can be staggering. Visual inspection of the data is no longer sufficient to make a meaningful interpretation of the information. Computer-driven solutions are required. For example, to analyze the DNA chip data, the discipline of bioinformatics—essentially a data mining exercise—emerged in the 1990s as a powerful melding of biology and computer science.

The formulas used in data mining are known as algorithms. Two common data mining algorithms are regression analysis and classification analysis. Regression analysis is used with numerical data (quantitative data). This analysis constructs a mathematical formula that describes the pattern of the data. The formula can be used to predict future behavior of data, and so is known as the predictive model of data mining.

Data that is not numerical (i.e., colors, names, opinions) is called qualitative data. To analyze this information, classification analysis is best. This model of data mining is also known as the descriptive model. The data mining process involves several steps such as,

- Defining the problem.
- Building the database.
- Examining the data.
- Preparing a model to be used to probe the data.
- Testing the model.
- Using the model.
- Putting the results into action.

7. Conclusion

Data Mining is one of the most important parameters in the age of digital era. The need and application of data mining has become essential to manage, organize, and disseminate information to the right users at right time. Though it is primarily intended for the business class, still then it has got practical implications in Libraries and Information Centers due to overwhelming growth of literature especially in digital formats. Now-a-days, more and more digital data are being collected, processed, managed and archived in Libraries and Information Centers to suit to the varied need of the user communities every day. Algorithms, software tools, and systems to mine it are critical to a wide variety of problems in all business, science, national defense, engineering, and health care including Libraries and Information Centers. Major steps involved in Data Mining are Data Cleaning where

data is to be cleaned from raw data for data mining and statistical modeling, Data Mart which is used to get ready with data for data mining, Derived Attributes on data, Modeling, Post Processing, and Deployment (www.answers.com/topic/data-mining). Data Mining thus has become indispensable for right management, organization and retrieval of data from ocean of information in Libraries and Information Centers.

References

1. Cullen D. (2005). Delving into data. Library Journal. 130(13), 30-32.
2. Gabriel Gomez, Data Mining Technology for Skills and Knowledge in Library Education at Chicago State University: Student Assessment in Live Text Software for Tracking and Teaching Data Mining. Prasad, (ARD) & Madalli (Devika P). Eds. International Conference on Semantic Web and Digital Libraries, 2007, Bangalore; ISI; 21-23 Feb. 2007, pp. 519-528.
3. Giudici (Paolo). (2005). Applied Data Mining Statistical Methods for Business and Industry. England; John Wiley & Sons; p.2.
4. http://en.wikipedia.org/wiki/Artificial_neural_network#Models (Accessed on 6.10.07)
5. http://en.wikipedia.org/wiki/Data_mining (Accessed on 6.10.07)
6. http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm (Accessed on 5.10.07)
7. http://www.answers.com/topic/data-mining?cat=biz-fin (Accessed on 5.10.07)
8. http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html (Accessed on 6.10.07)
9. http://www.rgrossman.com/dm.htm (Accessed on 10.10.07)

ABOUT AUTHOR

Dr. Rabinarayan Mishra is presently working as Lecturer in the Department of Library &Information Science, Mizoram University (A Central University), Aizawl, Mizoram. With regards to academic qualifications, Dr. Mishra has acquired M.A. in History, M.L.I.Science, L.L.B., and Ph.D., in Library & Information Science with PGDCA.