Relevance of Data Mining in Digital Library

R. N. Mishra and Aishwarya Mishra

***Abstract*—Data mining involves significant process of identifying the extraction of hidden predictive information from vast array of databases and it is an authoritative new technology with potentiality to facilitate the Libraries and Information Centers to focus on the most important information in their data warehouses. It is a viable tool to predict future trends and behaviors in the field of library and information service for deducing proactive, knowledge-driven decisions. Mechanized, prospective analyses of data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data Mining, Process of Data Mining, Knowledge Discovery in Databases, DBMS, Data Mining Techniques etc. etc. have been discussed in this paper.**

***Index Terms*—Data mining, KDD, artificial neural Networks, sequential pattern, modeling, DMT.**

I. Data Mining The Notion

Data Mining, defined in assorted ways with different implications by the experts, computer professionals and scientists relate to a heave of conglomerated data in multiple areas and it also can be matched with the term such as knowledge discovery. This infect involves a process of analyzing data from different perspectives to bring about user centric information that can be employed to increase revenue, costs, or both. Sporadic attempts have been done by the computer programmers to design data mining software where a number of analytical tools have been designed for analyzing data. It allows the users to analyze data not only in multiple dimensions and angles but also its categorization, and summarization of the relationships. 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. Evolution of data mining can be traced back when the business data were first stored in computers and technologies were generated to allow users to navigate through the data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation, to prospective and proactive information delivery and this evolutionary process is due to the support of three technologies such as, i) massive data collection, ii) high performance computing and iii) data mining algorithms (Pujari; 2001;p.44)

Manuscript received April 25, 2012; revised July 10, 2012.

R N Mishra is with Dept. of Lib. and Inf. Sciencce, Mizoram University, Aizawl, India (e-mail: rabinarayan_mishra@rediffmail.com),

Aishwarya Mishra is with Computer Science, SUIIT, Sambalpur University, Orissa, India (email: aishwaryamca@reiffmail.com),

Data Mining concept has been delimitated in multiple ways by different organizations, scientists etc. Wikipedia, has visualized data mining a non-trivial 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). Marketing Dictionary defines data mining as a process for extraction of customer information from a vast gamut of databases with the help of the feasible software to isolate and identify previously unknown patterns or trends. Multiple techniques with the help of technologies are employed in data mining for extraction of data from the heave of databases. Intelligence Encyclopedia has, however, defined data mining as a statistical analysis technique to retrieve useful data to ascertain trends or patterns (www.answers.com/topic/data-mining).

Data mining involves sorting through large amounts of data and picking out relevant information and though it is pertinent for the business organizations and financial analysts still then, Libraries and Information Centers can be excluded under its purview which can be increasingly used in the sciences to extract information from the enormous data sets generated by modern experimental and observational methods. Mention may be made that, Library and Information Service being one of the outstanding service field can well be accommodated in the arena of data mining as a good quantum of data are prevalent in the Libraries and Information Centers especially in digital environment.

Data Mining can be expressed in other multiple angles (Pujari; 2001; p.46) in library perspectives. It can be employed with library activities through;

Using assortment of techniques to identify nuggets of information or decision-making knowledge in the database and extracting these in such a way that they can be made use in other areas such as, decision support, prediction, forecasting and estimation. Discovering relation which connect variables in a database which can be interpreted for decision support system

Attaining by using pattern recognition techniques as well as statistical and mathematical technique for meaningful, new correlation patterns and trends by shifting through large amount of data stored in repositories s. In such a situation, the library and information center requires help from the subject experts and other outsourcing.

II. Data Mining- Phases of Development

Evolution of data mining has gone through different phases of development. Data mining initially emerged from

DOI: 10.7763/IJFCC.2012.V2.110

the business arena where the data were stored in computers and with the help of relevant technologies users tried to navigate in real time. Massive Data Collection, High Performance Computing and Data Mining algorithms are primarily the associated phenomena which on maturity coupled with high performance of relational database engines and broad data integrations precipitated to the employment of data mining technology. Data mining that identifies trends within data go beyond simple analysis, is a component of wider process known as 'Knowledge Discovery' from the databases which involves scientists from multiple arenas of disciplines, mathematicians, computer scientists and statiscians including the persons engaged in machine learning, artificial intelligence, information retrieval and pattern recognition (Pujari; 2001; p.2) . Through the use of sophisticated algorithms, users have the ability to identify key attributes of target opportunities.

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). Subsequent technological advances in data capture, processing power along with data transmission and storage capabilities, organizations like libraries and information centers in digital environment started integrating their various databases in to data warehouses for processing of information and retrieval from Centralized Data Management. 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.

III. Viability of Data Mining

Data Retrieval Technique also equally is a major component in the process like Data Mining that extracts data and information from archives and various national and international databases. Basically there are two prominent differences between the data retrieval and data mining which can be mentioned as follows.

Unlike data mining, the criteria for extracting information are decided beforehand so that they are exogenous from the extraction its-self.

Data Retrieval searches the relationships and associations between the unknown phenomena.

To discuss in nut-shell the historical perspectives of data mining, Wal-Mart named after Sam Walton of USA, one of the established corporation known as American Public Corporation was first to use the Data Mining Technology for transaction of business with relation to grocery and consumable where the technique was principally associated with many national and international agencies with a strong consumer focus such as, retail, financial, communication and marketing. This also enabled the agencies to determine the relationships among internal factors such as products, services, staffs and the external factors such as competition etc.

While information technology is employed to separate transaction and analytical systems, data mining links between the components. Software designed and applied in data mining analyzes various relationships and patterns. In the present ICT era market is flooded with multiple types of analytical softwares which can be utilized for data analysis, statistical inferences, machine learning and neural networks. It may not be out of place to mention that, the neural networks are based on the concept of Artificial Neural Network (ANN) which is associated with information processing paradigm inspired by the way of the biological nervous systems. Key elements involved in this paradigm is the novel structure of the information processing system which is composed of a large number of highly interconnected processing elements (neurons) 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 and information centers especially in a digital environment.

Classes: Library and information centers in the digital environment require accumulating pool of data to meet the versatile need of the predefined groups. Therefore, according to the type of clienteles such as, academicians, students, research scholars, engineers, doctors, and lawyers' etc. the library and information center process the digitally stored data for retrieval of information to meet demands of the users and the information so retrieved also could be used to increase traffic by having daily specials.

Clusters: Clustering is a method of grouping data into different groups according to logical relationship or consumer preference. It constitutes major class of data mining algorithms. Algorithms automatically partition the data space in to a set of regions or clusters. Clustering has manifold objectives such as, i) Uncovering natural groupings, ii) Initiating hypothesis for the data and iii) Locating consistent and valid organization of data.

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.

IV. Data Mining Process

Manifold activities are associated with data mining, as it

is an analytic process emphasizing more on developing a mechanism to explore data from a heave of data. Searching process of consistent patterns and/or systematic relationships between variables is involved in the process leading thereby, to validate the findings by applying the detected patterns to new subsets of data. Prediction is the ultimate target of the data mining which is otherwise known as Predictive Data Mining and this the most acceptable term of data mining having direct business applications. Predictive Data Mining is assorted with the process to identify data mining projects with an aspiration to identify a statistical or neural network model or set of models 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.

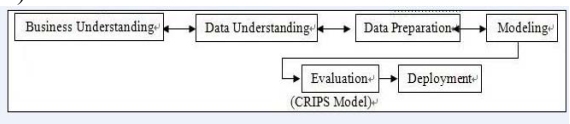
Concisely, data mining postulates three stages for exploration. The first stage of exploration normally is commensurate with data preparation involving in the cleaning process of data by removing the supererogatory elements resulting thereby, data transformations. The process is followed with selection of subsets of records where each case of data sets figures with large numbers of variables ("fields"). Depending upon the application of statistical methods, operation part is carried out to bring the number of variables to a manageable range. Depending on the nature of the problem, this first stage of the process of data mining can be applied anywhere between a simple choice of straightforward predictors for a regression model. This primarily leads to elaborate investigative analyses with the help of graphical and statistical methods for recognizing the suitable variables and complexities in the form of models for its application in the next stage. The second stage of the data mining operation is associated with data exploration where model building and validation are the key components and this process considers the application viability of various models to derive the predictive performance (i.e., explaining the variability in question and producing stable results across samples), which, however, is an elaborative process. Based on the competitive evaluation models, multiple techniques are available to achieve the target which is applied to the same data set for comparing the performance to choose the viable one. These techniques are considered as the core of predictive data mining which include: Bagging (Voting, Averaging), Boosting, Stacking (Stacked Generalizations), and Meta-Learning(http://www.statsoft.com /textbook/data-mining-techniques). The third and final stage relates to the deployment process that involves using the suitable model for operation in the above two phases to derive the consensus for generating predictions or probable result.

V. Models of Data Mining

Multiple models are in operations to carry out different function in data mining. However, this is outstanding in a outsized business oriented organization. Models primarily are meant to resolve problems presented in multi-tasking system. Need based model require developing to sort out problems and for emplacement of viable solutions to the emerging multiple problems in hefty organisations which could also be included in the purview of libraries where massive data are available.

A. CRISP Model

Cross-Industry Standard Process (CRISP) model in data mining was initiated by a European Consortium of Companies in 1990s primarily with an objective to serve as a non-proprietary standard process model. The model so designed by a group of companies basically works on four levels of steps mentioned below such as, Business Understanding; Data Understanding; Data Preparation;Modeling; Evaluation; and Deployment.(CRIPS Model)



CRIPS model emphasizes primarily to understand the objectives, missions of the organization where the relevant data obtained from various sources requires understanding to suit to the requirements of the organization. This is followed with data preparation from the accumulated data and with a suitable implementation of model it requires proper evaluation so as to discard the redundant data and the data so obtained through the process again requires meeting the target of the organization and at the end needs its operation for the growth and development of the organization.

B. Six Sigma Model

Another model has been developed on the platform of six sigma methodology which is further based on the logical relationship among data. In this model, data-driven methodology is applied for eliminating shortcomings and unwanted elements so that a qualitative product can be designed meeting the organization requirements and providing flawless services. This model has however, been identified by the developed countries like US as the suitable one and as such gained popularity. The model basically rests on five steps known as, DMAIC steps.

The model goes like this.

Define ⟶ Measure ⟶ Analyze ⟶ Improve ⟶ Control

The model stress upon defining a problem associated with the objectives of the mission and the course of action required to be adopted to overcome the problem. This can be done through proper analysis of the data before implementation and after distillation of data as per the need of the organization shall be implemented. The execution of data may or may not be suitable for operation by the end users who further opine he same for improvement and in such a situation, the organization takes all out efforts to disburse a qualitative and useful data or product to satisfy the requirements of the end users and finally the product is controlled. This is a cyclic process. Any new product or service needs to be passing through such a model so as to improve the quality and process control. Such type of model

is more prevalent in industries and service oriented organization.

C. SEMMA Model

Another model also which is in operative parallel to the Six Sigma Model as described is known as SEMMA model which was proposed by SAS Institute.

This model basically accentuates to the sample or population and this practically focus attention to the technical activities typically involved in a data mining project. The quantum of samples requires to be explored from the heave of data and thereafter, modification process is involved so as to eliminate the superfluous data not suitable to the organization objectives. The data thus obtained is to be routed through the suitable model of data mining to form the useful data for wider access by the end users. (Pujari; 2001; p.53).

VI. Elements of Data Mining

Multiple elements are associated with data mining. However, it 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.

Elements as stated above perform sequentially where, the data initially is extracted from massive of data or database relevant to the interested field which after transformation are loaded to the data warehouse for storing and management. The analyst with the help of devices applies the technology to reveal the useful parameters of the data for analysis and deliver the same in shape of useful information through a specific format to the end users for use. Data and or Information thus produced may be obtainable in shape of narrative or graph.

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. 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

VII. Implication of Data Mining Process in Library

Library practically in digital arena accumulates bulk of data for its users' communities which need to be processed to derive meaningful information as per the users' requirements. The process of deriving information has become more challenging tasks for the librarians and information officers who practically are not very much aware of technicalities involved in extraction process and to get redress the complications, the librarians including information centers take the expertise of computer personals working in the capacity of system administrators, data base managers etc. to retrieve data based on users' centric. This is prominent especially in a digital library system where multiple database both national and international use to operate and the library needs to extend the service round the clock. In the digital scenario consequent upon the application of information and communication technology in library services and to meet the versatile requirements of information by the user communities, the libraries and information centers started mounting up with more and more digital data which requires processing, managing, retrieval, disseminate and archive for future use. Due to proliferation of information wealth in the service segments like libraries and information centers, 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. To mention the implications of data mining in libraries, American Library Association website exhaustively has dealt with the issue. Data mining, according to ALA 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 has pointed out 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). Consequent upon the adoptability of Internet in library scenario, library could able to assemble massive data for its users in shape of e-resources. This is further aggravated with the consortia building and thus, the library could be recognized as a centre of massive data collection. This practically posed challenges for the librarians to retrieve useful data/ information, including its organization, management, dissemination and finally archiving. This has become essential as the professional related experts can understand, analyze the data as per the users' perspectives. (http://www.answers.com/topic/data-mining).

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.

VIII. Conclusion

Data Mining is one of the most imperative parameters in the age of digital era to implement in digital library scenario to extract the need based data/information from the heave of data accumulated in shape of electronic resources. E-resources refer to that kind of documents in digital formats which are made available to the library users through a computer based information retrieval system. E-resources in broad sense of the term include a variety of different publishing models, including online databases, sources from web pages, OPACs, e-journal articles, e-books, e-reports, e-databases, internet sources, print-on-demand (POD), electronic personal papers, e-mail messages, newsgroup postings, newsletters, government publications, electronic theses and dissertations, e-newspapers, CDs/DVDs, etc. In view of this, the need and application of data mining has become crucial 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 excavation relevant and useful data are critical to a redress assortment of problems in all business, science, national defense, engineering, and health care, railways including libraries and information centers especially in a digital environment.

References

[1] D. Cullen, "Delving into data." *Library Journal*, vol. 130, no. 13, pp. 30-32, 2005.

[2] G. 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," In. *International Conference on Semantic Web and Digital Libraries Eds. Prasad*, A. R. D. Madalli, D. P. Bangalore, DRTC; ISI; pp. 21-23 Feb. 2007, pp. 519-528.

[3] G. Paolo, "Applied Data Mining Statistical Methods for Business and Industry," England; John Wiley anf Sons. 2005, pp.2.

[4] A. K. Pujari, Data Mining Techniques. Hyderabad; University Press Pvt. Ltd., Inida, 2001, pp.44-60.

[5] Artificial neural network Models [Online]. Available: http://en.wikipedia.org/wiki/

[6] Data mining [Online]. Available: http://en.wikipedia.org/wiki/

[7] Data Mining: What is Data Mining? [Online]. Available: http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

[8] Data Mining. [Online]. Available: http://www.answers.com/topic/data-mining?cat=biz-fin

[9] Surprise [Online]. Available: http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html

[10] Rgrossman. [Online]. Available: http://www.rgrossman.com/dm.htm