Annals of Library and Information Studies
Vol. 58, March 2011, pp. 7-16

Usefulness and applications of data mining in extracting information from different perspectives

Jiban K Pal

Library, Documentation & Information Science Division, Indian Statistical Institute, 203, B. T. Road, Kolkata – 700108, Email: jiban@isical.ac.in

Discusses the concept of data mining that has emerged as a technique of discovering new meaning in data implicit in a large data warehouse, in order to facilitate better strategies and smart decisions. The standard tasks involved in data mining process are explained. Such pattern-seeking techniques usually performed with a wide range of related areas (viz. statistics, neural networks, genetic algorithms, machine learning, pattern recognition, knowledge-based systems, etc.) are described. Draws attention to useful applications of data mining in varying degrees and illustrates the benefit of this powerful technology with several instances of providing solutions to challenging problems. Also focuses on bibliomining opportunities to be useful to information retrieval, semantic analysis of unstructured texts, web-usage mining and to make proactive as well as knowledge-driven decision across library services; that help researchers in linking information for novel discovery and insight. The discussion illuminates the wrong things that happen with data mining applications and cautionary tales for avoiding serious consequences. Suggests use of data mining in combination with other techniques of evaluation, exploiting large data warehouses by skilled specialists, and advises for ethical uses without privacy invasion.

Introduction

With the advancement of technological innovations, various changes in approaches to organizing and retrieving information have been noticed. Taking advantage of these technological innovations information professionals have rendered many efforts in computerization, digitization, electronic access to information, data archiving, online analytical processing, etc. However, online processing and retrieval of information becomes a proactive area of library and information science. In fact, information professionals are trying to explore newer tools and techniques for knowledge management and discovery. Gradually the concept of 'data mining' has been evolved in modern information society as a technological solution to enhance the knowledge discovery in databases. It is even possible in distributed network environment. Today it defines the way to discover new meaning in data. Actually data mining performs "data processing using sophisticated data search capabilities and statistical algorithms to discover patterns and correlations in large preexisting databases"[1]. It can be utilized in any organization or system that needs to determine the patterns and relationships in their large dataset. The concept is equally important for business managers, policy makers, researchers, information specialists, data archivists, database developers, library managers, and seekers of information. Information analysts can provide a reasonable level of assurance to their results, commercially and otherwise useful through data mining efforts. Library and information professionals are experiencing this technique of knowledge discovery in a greater way by means of text-mining, biblio-mining, web-mining, etc. The idea becomes inevitable for discovering deeper information and enables categorization of data embedded in multiple source contents. In particular, data mining methodology extracts hidden predictive information from large databases by means of powerful technology with great potential to help the organizations (libraries) in providing most important information available in their data warehouses[2]. As such data mining allows one to find the needles hidden in one's haystacks of data, which makes it significant for library and information centers.

Data Mining – concept and definition

Data mining is considered to be an emerging technology that has made a revolutionary change in

the information world. The term 'data mining' (often called as knowledge discovery) refers to the process of analyzing data from different perspectives and summarizing it into useful information by means of a number of analytical tools and techniques, which in turn may be useful to increase the performance of a system. Technically, "data mining is the process of finding correlations or patterns among dozens of fields in large relational databases". Therefore, data mining consists of major functional elements that transform data onto data warehouse, manage data in a multidimensional database, facilitates data access to information professionals or analysts, analyze data using application tools and techniques, and meaningfully presents data to provide useful information. According to the Gartner Group, "data mining is the process of discovering meaningful new correlation patterns and trends by sifting through large amount of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques"[3]. Thus use of data mining technique has to be domain specific and depends on the area of application that requires a relevant as well as high quality data.

More precisely, data mining refers to the process of analyzing data in order to determine patterns and their relationships. It automates and simplifies the overall statistical process, from data source (s) to model application. Practically analytical techniques used in data mining include statistical methods and mathematical modeling. However, data mining and knowledge discovery is a rapidly growing area of research and application that builds on techniques and theories from many fields, including statistics, databases, pattern recognition, data visualization, data warehousing and OLAP, optimization, and high performance computing[4]. Worthy to mention that on-line analytical processing (OLAP) is quite different from data mining, though it provides a very good view of what is happening but can not predict what will happen in the future or why it is happening. In fact, blind applications of algorithms are not also data mining. In particular, "data mining is a user-centric interactive process that leverages analysis technologies and computing power, or a group of techniques that find relationships that have not previously been discovered"[5]. So, data mining can be considered as a convergence of three technologies – viz. increased computing power, improved data collection and management tools, and enhanced statistical algorithms.

Evolutionary steps in data mining

Data mining techniques are the result of a long process of research and have gone through various steps of evolution. Such evolution began when business data was first stored on computers and generated technologies to allow users for navigating their data in real time. Data mining algorithms have existed for at least ten years, but have recently been implemented as reliable and understandable tools. Now it is supported by further technologies that are sufficiently mature for navigation to prospective and proactive information delivery. In the evolution from business data to business information various steps have been noticed.

Table 1—Steps in the evolution of data mining (Source: www.thearling.com)

Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft.	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	SPSS, Comshare, Arbor, Cognos, Microstrategy,NCR	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases.	SPSS/Clementine, Lockheed, IBM, SGI, SAS, NCR, Oracle, numerous startups	Prospective, proactive information delivery

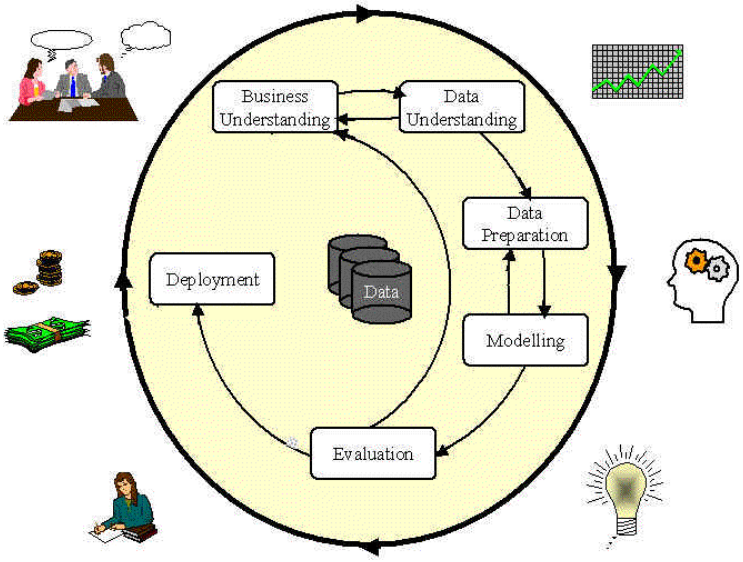

Fig.1—Tasks implicit in data mining process (Source: www.crisp-dm.org)

From the user's point of view, Squier[6] presented the four evolutionary steps in data mining (Table 1). These steps viz. data collection, data access, data warehousing, and decision making were revolutionary as they allowed new business questions to be answered accurately and quickly.

For decades major components of data mining technology have been under development in research areas such as statistics, artificial intelligence, and machine learning. But in recent times, the maturity of these techniques coupled with high-performance relational database engines and broad data integration efforts make these technologies more effective for current data warehouse environments[7].

Data mining process

Primarily, the data mining requires a standard process, data store or warehouse, technologies and expertise. A data mining process must be reliable and repeatable by people with little data mining skills. However, a standard data mining process should involve the following tasks

1. *Job understanding* – determines the job objectives, job background, data-mining objective, situation assessment, risk and controversies, success criteria, etc.
2. *Data understanding* – collects initial data, describe data, explore data, and verify the data quality.
3. *Data preparation* – involves data set description, data selection, assessment, consolidation, cleaning, derive data attributes, data formatting, etc.
4. *Process modeling* – identifies modeling techniques based on data mining objectives, parameter setting and generate test designs, model assessments, etc.
5. *Process evaluation* – evaluates results, review the process, approve the model, and determine the forward steps for possible actions or decisions.
6. *Deployment* – plans for monitoring and maintenance, produce final reports, reviews experiences, presents documentation, etc.

Figure 1 illustrates the data mining process model (based on CRISP-DM) in a large data warehouse.

In fact, data mining was broadly involves in the interpretation and validation of discovered patterns to ensure the real information being discovered for knowledge representation. Extensively it uses the techniques from machine learning, statistics, pattern discovery algorithms, and many other fields. In data mining, most commonly used techniques are – decision trees, nearest neighbor classification, neural networks, rule induction, K-means clustering, genetic

algorithms, bayesian networks, etc.[8] Among these techniques decision tree models are relatively fast and easy to understand; but neural networks (although powerful modeling tool) are relatively difficult to interpret (compared to rule-induction, decision trees, sequential patterns, etc.) and also require significant amounts of time. Many of these techniques have been in use for more than a decade in specialized analysis tools and now evolving to integrate directly with industry-standard data warehouse and OLAP platforms. CRoss Industry Standard Process for Data Mining (CRISP-DM) is a popular initiative (launched in 1996) becomes useful to data mining vendors, system suppliers, business consultants, and end-users. As the name suggests, this models is non-proprietary, industry natural, tool natural and provides experience based uniform framework for guidance as well as flexible enough to account for differences.

Useful applications of data mining

Data mining was originally developed to act as expert systems to solve problems and did not require assumptions to be made about data. Now it has little interest in the mechanics of the technique but requires understanding of data and business problem to predict the trends and behaviours in an automated way. However, data mining has become useful to sort out previously unknown patterns in a large data system, though often used for data dredging and data snooping - refers to the practice of misusing data mining techniques to show misleading or false conclusions. David Jensen has clearly noted the darker side of data mining regularly arised in knowledge discovery applications and also devised a variety of approaches to avoid such pitfalls in data mining practices[9]. Worthy to mention that dredging and snooping can be used as exploratory tools when developing and clarifying hypotheses.

Broadly, the data mining could be useful to answer the queries on –

- Forecasting with regard to what may happen in the future.
- Classifying things into groups by recognizing patterns.
- Associating similar events that are likely to occur together.
- Clustering the peoples into groups based on their attributes.
- Making the sequence what events are likely to lead to whom.

However, it can be used in a predictive manner for a variety of applications in order to achieve organizational goal. Common uses of data mining are given below:

a) *Fraud or non-compliance anomaly detection* – Data mining isolates the factors that lead to fraud, waste and abuse. The process of compliance monitoring for anomaly detection (CMAD) involves a primary monitoring system comparing some predetermined conditions of acceptance with the actual data or event. If any variance is detected (an anomaly) by the primary monitoring system then an exception report or alert is produced, identifying the specific variance. For instance credit card fraud detection monitoring, privacy compliance monitoring, and target auditing or investigative efforts can be done more effectively[10].

b) *Intrusion detection* – It is a passive approach to security as it monitors information systems and raises alarms when security violations are detected. This process monitors and analyzes the events occurring in a computer system in order to detect signs of security problems. Intrusion detection systems (IDSs) may be either host based or network based, according to the kind of input information they analyze[11]. Over the last few years, increasing number of research projects (MADAM-ID, ADAM, Clustering project, etc.) have been applied data mining approaches (either host based or network based) to various problems (construction of operational IDSs, clustering audit log records, etc.) of intrusion detection[12].

c) *Lie detection (SAS Text Miner)* – SAS institute introduced lie-detecting software, called SAS Text Miner. Using intelligence of this tool, managers can be able to detect automatically when email or web information contains lies. Here data mining can be applied successfully to identify uncertainty in a deal or angry customers and also have many other potential applications[13]. Many other market mining tools are also available in real practice viz. Clementine, IBM's Intelligent Miner, SGI's MineSet, SAS's Enterprise Miner, but all pretty much the same set of tools.

d) *Market basket analysis (MBA)* – Basically it applies data mining technique in understanding what items are likely to be purchased together according to association rules, primarily with the aim of identifying cross-selling opportunities. Sometimes it is also referred to as product affinity analysis. MBA gives clues as to what a customer might have bought if an idea had occurred to them. So, it can be used in deciding the location and promotion of goods by means of combo-package and also can be applied to the areas like analysis of telephone calling patterns, identification of fraudulent medical insurance claims, etc.[14].

e) *Aid to marketing or retailing* – Data mining could help direct marketers by providing useful and accurate trends on purchasing behavior of their customers and also help them in predicting which products their customers may be interested in buying. In addition, trends explored by data mining help retail-store managers to arrange shelves, stock certain items, or provide a certain discount that will attract their customers. In fact data mining allows companies to identify their best customers, attract customers, aware customers via mail marketing, and maximize profitability by means of identifying profitable customers[15].

f) *Customer segmentation and targeted marketing* – Data mining can be used in grouping or clustering customers based on the behaviors (like payment history, etc.), which in turn helps in customer relationship management (epiphany) and performs targeted marketing. Usually it becomes useful to define similar customers in a cluster, holding on good customers, weeding out bad customers, identify likely responders for business promotions.

g) *Phenomena of "beer and baby diapers"* – This story of using data mining to find a relation between beer and diapers is told, retold and added to like any other legend. The explanation goes that when fathers are sent out on an errand to buy diapers, they often purchase a six-pack of their favorite beer as a reward. An article in The Financial Times of London (Feb. 7, 1996) stated, "The oft-quoted example of what data mining can achieve is the case of a large US supermarket chain which discovered a strong association for many customers between a brand of babies nappies (diapers) and a brand of beer[16].

h) *Financial, banking and credit or risk scoring* - Data mining can assist financial institutions in various ways, such as credit reporting, credit rating, loan or credit card approval by predicting good customers, risk on sanctioning loan, mode of service delivery and customer retention (i.e. build profiles of customers likely to use which services), and many others. A credit card company can leverage its vast warehouse of customer transaction data to identify customers most likely to be interested in a new credit product. In addition, data mining can also assist credit card issuers in detecting potentially fraudulent credit card transaction. In general, data mining methods such as neural networks and decision trees can be a useful addition to the techniques available to the financial analyst[17].

i) *Medicare and health care* – Applying data mining techniques, it is possible to find relationship between diseases, effectiveness of treatments, to identify new drugs, market activities in drug delivery services, etc. However, a pharmaceutical company can analyze its recent sales to improve targeting of high-value physicians and determine which marketing activities will have the greatest impact in the next few months. The data needs to include competitor market activity as well as information about the local health care systems. Such dynamic analysis of the data warehouse allows best practices from throughout the organization to be applied in specific sale situation.

j) *Corporate surveillance* – Implies the monitoring of a person or group's behavior by a corporation, which is highly possible through data mining process. So it can be used as a form of business intelligence that enables the corporation to better tailor their products and services to be desirable by their customers. Normally the organizations that have enemies who wish to gather information about the group members or activities face the issue of infiltration usually followed a process of data mining. Thus, surveilling party may put pressure on certain members of the target organization to act as informants i.e. disclose the

information they hold on the organization and its members.

k) *Scientific enquery and research analysis* - Data mining can be applied in diverse areas of scientific activities and allows researchers to perform and analyze scientific data more effectively. It flourishes in telecommunication industry due to adequate amount of data available in structured form, which helps in network analysis, tracking unusual groups of customers, launching profitable service vouchers, etc. For instance, "if a customer makes more than X calls per hour to country Y, then apply treatment Z" – thus data mining (techniques) can be applied to all customers based on these (X, Y, Z) parameters. Many scientific enquiries in atmospheric science, biomedical engineering, information retrieval, patent analysis, remote sensing, etc. prefer to use data mining for novel discovery.

l) *Criminal investigation and homeland security* - Data mining becomes useful in suspecting criminals as well as apprehending these criminals by examining trends in location, crime type, habit, and other patterns of behaviors. It also very useful to track fraud in order to increase the safety and security of a country.

Data mining can also be applied successfully in several occasions like – parts failure detection web site personalization[18], bio-informatics[19], text mining, text clustering, web mining[20], and many others. A supermarket becomes an information broker and basketball teams use data mining to track game strategy. Web mining finds the affinity of visitor to web pages analyzing web logs, identifies popular pages, analyse links that are hard to find, etc. However, data mining assist researchers by speeding up their data analyzing process, thus allowing them more time to work on other projects.

Data mining in libraries (or Bibliomining)

The concept of data mining has occasionally discussed in traditional library settings since late 1990's[21]. Over the past few years, library professionals have been trying for the integration of data mining in different library and information services, so called "bibliomining" as coined by Nicholson and Stanton in 2003[22]. Bibliomining is derived from the terms bibliometrics and user-based data mining, which brings together through data warehousing in various library services. Combining these two data sources into one data warehouse allows researchers and library managers to more fully understand the information space they have created[23]. In addition, a properly designed data warehouse enables evaluation of a service for gaining understanding of library use across the services. So, the librarian's need for decision-making and the scholars' needs for research should drive the fields captured in data warehouse. So, the concept becomes useful to understand patterns of behavior among library users and patterns of information resource use throughout the institution. For instance, data mining that would be useful in digital library evaluation is web-usage mining. Here data mining extensively uses the transaction log records from a web server to discover the patterns by means of efficient techniques, followed by an interpretation and validation of discovered information pertaining to track the usage pattern (or such other) of digital library services[24].

However, the basic objective of bibliomining is to explore the dataset for patterns that are novel and useful for libraries. It is basically the combination of data mining, bibliometrics, statistics, and reporting tools used to extract patterns of behavior-based artifacts from library systems. Integrating bibliomining into current research and evaluation efforts will allow the library managers and researchers a more complete idea of the resources contained in their library and how the resources are being accessed by the users[25]. Each bibliomining analysis can reveal a pattern of activity within the library and these patterns may have potential benefits in three different levels;

a) benefits for individuals through improved library services,

b) benefits for library management through the provision of improved decision-making information, and

c) benefits for the institution that the library serves through reporting of relevant patterns of user behavior.

Therefore, by providing information on the performance and utility of the library as a unit,

bibliomining can provide justification for continued financial and institutional support for library operations. These three levels further describe the bibliomining opportunities and applications in libraries.

But, bibliomining is one of the techniques to be used in combination with other techniques of evaluation, such as LIBQUAL[26], cost-benefit analyses, surveys, etc. While the information provided through bibliomining is useful, it needs to be supplemented by more user-based approaches to provide a more complete picture of the library system. It is also important to ensure the authenticities in record creation and acquisition of appropriate data as well as other interactions with the system. This consideration argues on the formation of data sharing consortium among the libraries of similar nature is needed. In fact, the bibliomining works optimally on very large data sets. There is no doubt that bibliomining can provide deeper understanding of the resources, services, and users of a library system in order to aid library decision-makers for discovering meaningful information on the access and utilization of their existing services. Above all, data mining technology is highly useful to information retrieval, semantic analysis, navigation of unstructured texts, digital library evaluation, web-usage mining and to make proactive as well as knowledge-driven decision across library services. Thus helps library managers in extracting meaningful information and researchers in linking information to knowledge discovery.

Success stories of data mining application

Data mining technology encompasses a rich collection of proven techniques that is both commercial and noncommercial realms. In commercial realm (industry applications) multiple techniques are used. Say, cluster detection technique can be applied to identify the clusters of customers, may be followed by a predictive algorithm applied to some of the identified clusters to discover the expected behaviors of the customers in those clusters. In fact data mining can be successfully applied to a variety of business purposes (viz. customer segmentation, market basket analysis, risk management, delinquency tracking, demand prediction, etc.); and in the practical areas like credit card analysis, insurance claims fraud analysis, transport logistics management, telecommunication call record analysis, consumer goods promotion analysis, etc. Noncommercial use of data mining is strong and pervasive in the research area. In oil exploration research, data mining techniques discover locations suitable for drilling based on potential mineral and oil deposits. In medical research this technology helps researchers with discoveries of correlations between diseases and patient characteristics. The scientific community makes use of data mining to a moderate extent, but the technology has widespread applications in marker research. Though many companies have deployed successful applications of data mining, but this technology can be reasonably beneficial to any corporations, governments, society and individual person who is looking to leverage a large data warehouse for better strategies to best reach them. In reality, data mining has done for many instances as given in the following paragraphs.

a) An initiative that has attracted significant attention for the application of data mining is Terrorism Information Awareness (TIA) project in May 2003; conducted by the Defense Advance Research Project Agency (DARPA), immediately after the terrorist attacks on September 11, 2001. TIA research project was primarily designed to develop technologies for the detection of terrorist groups planning attacks against the US government (both inside and outside the country), and to create experimental prototype tools. This project potentially improved data mining capabilities of several agencies by means of adopting new technologies like automated rapid language translation, enhanced collaboration or decision support tools, and improved search and pattern recognition. These technologies enabled more comprehensive and thorough mining of transactional data, such as immigration records, passport and visa applications, car rentals, driving license renewals, airline ticket purchases, criminal records, etc.[27].

b) Major US banks applied data mining for customer attrition prediction. First they segmented the customers based on financial behavior and build attrition models for three different segments. Thus 40-50% of attritions were predicted. Those banks also targeted credit marketing based on credit balances (12-15 months) of the segmented

customers, and increased customer responses about four-times applying new response model.

c) Network intrusion detection using a combination of sequential rule discovery and classification tree techniques on 4 GB DARPA data won over knowledge engineering approach. A total of 4 gigabytes data was gathered from a wide variety of intrusions simulated in a military network environment of DARPA over a period of seven weeks. In order to detect intrusions link analysis and sequential analysis methods were employed to determine the statistical attributes of network connections of profile records. The JAM (Java Agents for Meta-learning) project of Columbia University provides detailed description of the entire process[28].

d) HSBC successfully applied data mining in satisfying the need to cross-sell more effectively by identifying profiles that would be interested in higher yielding investments, and reduced direct mail costs by 30% while garnering 95% of the campaign's revenue[29].

e) It has been reported that data mining helped the federal government to recover millions of dollars in fraudulent Medicare payments[30]. US internal revenue service also needed this technology to improve customer service and scheduled its workforce to provide faster and more accurate answers to questions.

f) Data mining successfully applied to a variety of scientific problems emerged in biomedical engineering (DNA sequencing, sorting gene expression, gene therapy, etc.), remote sensing (on a satellite), information retrieval (semantic analysis of texts, cross-lingual text retrieval, clustering information repositories, etc.), and so many. Many scientific organizations in India and abroad have been employed data mining technology to deliver powerful research and analysis tools for bio-informatics, remote sensing, medical imaging, information retrieval, atmospheric, and astronomical research[31].

g) The US Drug Enforcement Agency needed data mining to be more effective in their drug "busts", and analyzed suspects cell phone usage to focus investigations.

h) In fact, US government regularly uses data mining to track fraud in order to increase security[32].

Limitations in adopting data mining

From the utilitarian view, data mining is extremely good because it enables corporations to minimize risk and increases profit, helps government to strengthen security system, and benefit the society by speeding up the technological advancement[33]. However, it increases the risk of privacy invasion, a major way that companies can invade the consumers' privacy, and any department can disclose or misuse the personal information. In fact, if the government wants to create laws in order to protect the people's privacy right, still it is extremely difficult to monitor the disparate resources (viz. servers, databases, web sites, blogs, etc, as they are globally accessible); thus creating tremendous difficulty for the government to enforce the laws on data smuggling. While data mining products can be very powerful tools, they are not self-sufficient applications. Successful data mining requires skilled technical and analytical specialists who can structure the analysis and interpret the output that is created[34]. In fact such pattern seeking tools cannot reveal everything needed to make smart decisions, i.e., data mining techniques are not hundred percent accurate, thus mistakes do happen which can have serious consequences[35]. Basic constraints that affect the development of data mining systems are – unavailability of large data sets with adequate density of the relevant data to satisfy the real interest, lack of efficient mining algorithms to support universal data mining capability, etc. However, diverse sets of data available in wide variety of formats and temporal characteristics would require either various mining systems for individual data sets or universally accepted data format with a single data mining system. Otherwise, machine-readable metadata with certain standard in order to support data mining capability could bring a reasonable solution in this direction[36]. Importantly, using only data mining (bibliomining) to understand a library or information system may not be proper or complete. Simply it is one of the tools to be used in combination with other techniques of library evaluation. Thus, it is essentially require for predictive decisions in library situation, but not to be considered to make an ultimate decision.

Conclusion

Data mining is a way to discover new meaning in data, performs data processing using sophisticated data search capabilities and statistical algorithms, which can be utilized in any organization or system that needs to determine the patterns or relationships implicit in a large data warehouse for better strategies to best reach them. It can be reasonably beneficial to any corporate industries, financial institutions, retailers, pharmaceutical firms, security agencies, government departments, online service providers, libraries, and individual researchers too. It can be used for a variety of applications in both public and private sectors. Corporate industries and financial institutions often use data mining to increase sales, reduce costs, improve market performance, enhance customer base by means of developing models for credit scoring, risk assessment, fraud detection, etc. Retailers can use data mining to assess the effectiveness of product selection, placement decisions, coupon offers, and products that are often purchased together. Pharmaceutical firm frequently uses data mining of chemical compounds and genetic materials to guide the research on new medicines or treatments for diseases. Internet and telephone service providers sometimes apply data mining techniques to assess the subscribers and strength of the competitors. Recently data mining have been increasingly used in public sectors for many purposes viz. detecting fraud and waste, measuring and improving program performance, identifying fraudulent claims for payment, adjusting resource allotments, predicting crime patterns and locations, identifying terrorist activities, tracking individual terrorists, controlling aviation traffics, etc. In no doubt data mining becomes an essential tool for homeland safety and security, marketing, process control, manufacturing, network detection, and many others. Information analysts can provide a reasonable level of assurance to their results for commercially and otherwise viable through data mining efforts. Libraries are gradually realizing the value of creating a large data warehouse that can be mined aiming in view to strengthen library decision-making for diverse operations. Moreover data mining allows researchers to collect and analyze data more efficiently for new medications, biological agents, gene therapy, etc. to cure deadly diseases like cancers, AIDS, etc. and aids large projects (Human Genome project) so easily and quickly[37]. So, data mining tools and techniques can be useful to several activities to make proactive and knowledge-driven decisions. Data mining should be used in an ethical way. This discussion would be helpful to library managers and researchers in putting the bibliomining in its proper context.

Acknowledgements

The author gratefully thanks Dr. Usha Mujoo Munshi (Librarian, Indian Institute of Public Administration, New Delhi) for inspiration and valuable comments on the draft of this paper. He also wishes to thank all of his beloved colleagues for providing their constant support in preparing this paper.

References

1. The Free Dictionary, by Farlex Inc. © 2010, Available at: http://www.thefreedictionary.com/data+mining (accessed on 6th January 2010).
2. Folorunso O and Ogunde A O, Data Mining as a technique for knowledge management in business process redesign, *The Electronic Journal of Knowledge Management*, 2 (1) (2004) 33-44, Available online at: www.ejkm.com (accessed on 8th January 2010).
3. Larose D T, *Discovering knowledge in data: an introduction to data mining*, (John Wiley; New York), ©2005, Available also at: http://www.dataminingconsultant.com/DKD_Larose_Preface.htm (accessed on 10th January 2010).
4. Klosgen W and Zytkow J M (eds.), *Handbook of data mining and knowledge discovery*, (OUP; Oxford), 2002.
5. Kantardzic M, *Data mining: concepts, models, methods, and algorithms*, (John Wiley; New Jersey), 2003.
6. Squier L, What is data mining, Available at: http://www.dama-ncr.org/Library/2001.11.14-Laura%20Squier.ppt (accessed on 12th January 2010).
7. Thearling K, Information about data mining and analytic technologies (white paper), Available at: http://www.thearling.com/index.htm (accessed on 16th January 2010).
8. Han J and Kamber M, *Data mining: concepts and techniques*, (Morgan-Kaufman; New York), 2001.
9. Jensen D, Data snooping, dredging and fishing: the dark side of data mining, *SIGKDD Explorations*, 1 (2) (January 2000) 52-54.
10. Goldschmidt P S, *Compliance monitoring for anomaly detection*, Patent no. US 6983266 B1, issue date January 3, 2006, Available at: www.freepatentsonline.com/6983266.html (accessed on 16th January 2010).
11. Bace R, *Intrusion Detection*, (Macmillan Technical Publishing), 2000.
12. Smyth P, Breaking out of the Black-Box: research challenges in data mining, Paper presented at the Sixth Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD-2001), held on May 20 (2001), Santra Barbara, California, USA.
13. SAS Institute Inc., Lie detector software: SAS Text Miner

(product announcement), *Information Age Magazine*, [London, UK], February 10 (2002), Available at: http://www.sas.com/solutions/fraud/index.html (accessed on 22nd January 2010).

14. Berry M J A and Linoff G S, *Data mining techniques: for marketing, sales, and relationship management*, 2nd edn (John Wiley; New York), 2004.
15. Delmater R and Hancock M, *Data mining explained: a manager's guide to customer-centric business intelligence*, (Digital Press, Boston), 2002.
16. Fuchs G, Data Mining: if only it really were about Beer and Diapers, *Information Management Online*, July 1, (2004), Available at: http://www.information-management.com/news/1006133-1.html (accessed on 24th January 2010).
17. Langdell S, Use of data mining in financial applications, (Data Analysis and Visualization Group at NAG Ltd.), Available at: http://www.nag.co.uk/IndustryArticles/DMinFinancialApps.pdf (accessed on 26th January 2010).
18. Mobasher B, Dai H, Luo T and Nakagawa M, Discovery and evaluation of aggregate usage profiles for Web personalization, *Data Mining and Knowledge Discovery*, 6 (1) (2002) 61-82.
19. Wang J T L, Zaki M J, Toivonen H T T and Shasha D (eds.), *Data mining in bioinformatics*, (Springer-Verlag, London), 2005.
20. Chakrabarti S, *Mining the web: discovering knowledge from hypertext data*, (Morgan Kaufmann; Amsterdam), 2003.
21. Banerjee K, Is data mining right for your library? *Computers in Libraries*, 18 (10) (1998) 28-31.
22. Nicholson S and Stanton J, Gaining strategic advantage through bibliomining: data mining for management decisions in corporate, special, digital, and traditional libraries, In *Organizational data mining: Leveraging enterprise data resources for optimal performance*, edited by H. Nemati & C. Barko, (Idea Group Publishing; Hershey, PA), 2003, p.247-262.
23. Nicholson S, Preprint of the basis for bibliomining: frameworks for bringing together usage-based data mining and bibliometrics through data warehousing in digital library services, *Information Processing and Management*, 42 (3) (2006) 785-804.
24. Srivastava J, Cooley R, Deshpande M and Tan Pang-Ning, Web usage mining: discovery and applications of usage patterns from Web data, *SIGKDD Explorations*, 1 (2) (2000) 12-23.
25. Liu B, *Web data mining: exploring hyperlinks, contents, and usage data*, (Springer-Verlag; Berlin), 2007.
26. LIBQUAL – Charting Library Service Quality (statistics and measurement program), (Association of Research Libraries: © 2009), Available at: http://www.libqual.org/home (accessed on 24th January 2010).
27. Department of Defence (Govt. of US), Report to Congress regarding the Terrorism Information Awareness Program: Executive summary, 20 May 2003, p.1-2.
28. JAM Project, Columbia University, New York, Available at: http://sneakers.cs.columbia.edu/JAM/PROJECT/ (accessed on 12th February, 2010).
29. Squier L, What is data mining (slide no. 32 & 33), hosted by Data Management Association – National Capital Region (DAMA-NCR), Available at: http://www.dama-ncr.org/Library/2001.11.14-Laura%20Squier.ppt (accessed on 12th February 2010).
30. George C, Data mining taps and trends, *Government Executive Magazine*, October 1 (2000), Available at: http://www.govexec.com/tech/articles/1000managetech.htm (accessed on 16th February 2010).
31. Grossman R. L, Kamath C, Kegelmeyer P, Kumar V and Novnburu R. (eds.), *Data mining for scientific and engineering applications*, [Massive computing series, vol.2], (Springer-Verlag; London), 2001.
32. Seifert J W, Data mining and homeland security: an overview, *CRS report for congress*, January 18 (2007), Available at: http://www.fas.org/sgp/crs/intel/RL31798.pdf (accessed on 24th February, 2010).
33. Chhay H, Data mining, Coen288 – Software Ethics (Santa Clara University), Available at: http://cseserv.engr.scu.edu/StudentWebPages/hchhay/hchhay_FinalPaper.htm (accessed on 26th January 2010).
34. Seifert J W, Data mining: an overview, *CRS report for congress*, December 16 (2004), Available at: http://www.fas.org/irp/crs/RL31798.pdf (accessed on 22nd February, 2010).
35. McDonough B, Developers of deep Web mining technology acknowledge concerns about how personal data is used, but say the problem must be handled by those who decide what information goes on the Internet. NewFactor Sci:: Tech 2002; 5.
36. Behnke J and Dobinson E, NASA workshop on issues in the application of data mining to scientific data (summary), *ACM SIGKDD Explorations Newsletter*, 2 (1) (June 2000) 70-79.
37. Dubitzky W, Granzow M and Berrar D (eds.), *Fundamentals of data mining in genomics and proteomics*, (Springer-Verlag, New York), 2007.