DESIDOC Journal of Library & Information Technology, Vol. 34, No. 5, September 2014, pp. 401-406 DOI: 10.14429/djlit.34.6722

Data Mining: Competitive Tool to Digital Library

Tariq Ahmad Lone and Rafi Ahmad Khan

The Business School, University of Kashmir, Srinagar-190 006
E-mail: tariq380@gmail.com, mca_rafi@yahoo.co.in

ABSTRACT

The increasing popularity of world wide web is providing an easy and quick alternative to get the required information through the internet which has threatened the existence of traditional libraries. The human capital associated with the management of libraries in the contemporary world is facing the challenge of crafting strategies to continuously augment their services with the help of information technology to provide quality customised services to the end users. To address this challenge libraries can use data mining to extract strategic information from the data captured in the libraries' data ases. The capability of data mining to solve some of the vital problems has encouraged libraries to implement this intelligent tool of data mining. In this paper, an attempt has been made to educate stakeholders on the issues and applications of data mining in libraries.

Keywords: Data mining, information technology, extraction, transformation and loading, Knowledge discovery

1. INTRODUCTION

During the pre-web days, libraries were one of the main sources for individuals seeking information. With the increasing demand for information, the focus of the libraries has been to build large collections and repositories for knowledge in the form of print media[1]. The innovations in IT have made the accessibility and storage of information so much easier and faster that the demand for more, specific and reliable information by users has caused the libraries to introduce new IT systems in the transmission and preservation of their resources. The implementation of new information technologies in the libraries has improved their functions and the services provided to their subscribers. Libraries today are not just facilities for printed resources or places to sit and study, but have extended their services beyond the physical walls of the libraries to provide access to information technology media[2].

The growth and exceeding popularity of the internet has made the world wide web become a new concept of sharing information. The world wide web has enabled information to be obtained and shared quickly and easily throughout the world in a secured environment. It has revolutionised the information services by providing an alternative way for individuals to get the information they require. The exceedingly popular usage of it by individuals has imposed a threat to the existence of libraries as the main source for getting information. It could cause the libraries' patrons to move away from libraries to seek the information they require from the web instead. Libraries' management need to plan strategically ways of enhancing and introducing new services with the intention of encouraging their patrons to use their libraries to obtain information. Libraries' management need to maintain their patrons' confidence to seek their libraries as a reliable source of information[1]. In fact, information professionals are trying to explore newer tools and techniques for knowledge management and discovery[3].

One of the strategies of the libraries' management was to host their own library website that would provide information about their collections and services to their patrons, within and outside the premises of the library[1]. The libraries' websites should be able to provide its patrons expedite access to the information they seek; help to locate the answers required quickly; and at the same time, promote the awareness of their library to the public[4]. The implementation of the libraries' websites and the use of new IT has, indirectly, forced librarians to become more computer literate and to acquire knowledge about internet technologies which are used to plan, design and implement the services required by their patrons[5].

2. LIBRARY SYSTEMS

Computers were used to overcome the difficulty of retrieving documentation as early as in the 1950's, where files were stored on magnetic tapes and ran on mainframe computer systems[6]. In the 1970s,

Received 11 March 2014, Revised 21 July 2014, Online published 4 September 2014

computers began to be used more widely, where the mainframe computer was capable of storing data in hard disk and there was an increase of processing power, memory and performance. Then in the late 1970s and early 1980s, there was the introduction of 'online search', where information could be accessed online[6]. Libraries began to implement the information communication technology (ICT) with the introduction of online search. This system was to perform most of the functions in the libraries, which included collection assessments, acquisition, cataloguing, and circulation information[7]. As large corpora of data accumulate, automated and semi-automated analysis of their contents reveals patterns and allows establishment of fact patterns invisible to the naked eye. Libraries and librarians have an exciting opportunity to support this technology-based work[8]. The integrated library system consists of the following:

2.1 Cataloguing

The Online Public Access Catalogue (OPAC) is one of the common library information systems where all the resources or information of all materials and its' location in the library are stored. The OPAC contains information of the materials such as the title, authorship, publication, physical description, access method, unique call number, subject headings and other necessary information[7].

2.2 Information Acquisition

The acquisition information system contains information about items ordered from vendors, tracking of the orders, delivery and processing dates of the items by the library[8]. These acquisition systems do not have some of the important features such as the ability to:

(a) Predict the time taken to purchase the materials,

(b) Predict the cost of resources in the future,

(c) Track the vendors' performance, such as delivery of the ordered materials as scheduled, and

(d) Automatically track missing purchases or late delivery of materials.

Without these features in their system, the libraries would have difficulty in tracking and monitoring their acquisition especially, with the increasing amount of materials demanded by the patrons and the need to increase the resources in the libraries.

2.3 Search

The library database captures and stores patrons' information. This information contains the clients' name, address, contact information and other necessary information[7]. Often the library management needs to understand their clients' to serve them better, and to perform analysis on the database to analyse their clients' behaviour. This type of analysis is difficult to be performed by the librarian, because it would require specific skills and analysis tools, which are not integrated into this system.

2.4 Information Circulation

Circulation of information is the information related to the resources of the library, such as the checkout date, due date, call number and some basic information of the borrower or the user of the resources. The information circulation also captures information of resource materials that are requested by the clients' or reserved by the librarian but these records are erased when the materials are returned back to the library[7]. This circulation information should have some of the following features, which are useful for the library management:

(a) Ability to produce information on the borrowing trend of books or resources

(b) Ability to determine the relationship between the patrons and the materials

(c) Ability to list materials, which are in demand or being borrowed very frequently, and

(d) Ability to list the books or materials, which are not being borrowed or used at all by the patrons.

These features are very useful for the library management, because it could help the libraries to:

(a) Justify the purchasing of books or resources for the library,

(b) Prepare budget reports – where these information can be used to justify or support their reports,

(c) Improve their services by increasing the resources which are in demand, and

(d) Remove materials or resources which are not being used.

3. CHALLENGES POSED BY IT

The IT innovations have posed the following challenges to libraries[5]:

(a) Libraries have had to implement a virtual method for information access

(b) Libraries have had to keep up with the technology enhancement which increase their budget

(c) Libraries' staff needed training on the use of these new technologies, including maintenance ensuring the availability of increasing information

(d) Libraries ought to ensure their resources or collections are being fully utilised

(e) Libraries that have chosen to adopt the new technologies did not have the resources and the expertise to maintain these systems, and

(f) Libraries' online system has had to be accessible all the time to provide digital information continuously.

To overcome some of these challenges and issues the libraries' management requires strategic information to make decisions that would enable them to establish their goals, objectives and monitor the results. Ponniah[9] states that strategic information has the following characteristics:

(a) Integrated–a single and complete view of the organisation.

(b) Data Integrity–information has to be accurate and in line with organisational objectives.

(c) Accessible–easily obtained whenever required for analysis.

(d) Credible–the information needs to be credible.

(e) Timely–the information has to be available within the required time.

Strategic information, previously not easily available, has been facilitated with the advent of a new paradigm in IT called Data mining. Data mMining is specific to produce information based on the pattern(s) generated by the data captured in the systems. The libraries require strategic information to improve the services they provide, as a result, libraries should implement data mining systems to obtain the required information. With the improvement and introduction of new services to their patron, the libraries can still be one of the main sources of information.

4. DATA MINING

Data mining is defined as 'The automatic extraction of useful, often previously unknown information from large databases or data sets'[10]. One of the popular definitions for data mining is the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data to discover meaningful patterns and rules[11]. This can be simplified as a process of getting information or knowledge from large amounts of data. That is why sometimes data mining is also known as 'knowledge mining', a synonym for another popularly used term known as, 'knowledge discovery in databases (KDD)'[12].

Data mining may be described as a process of discovering interesting information from large amounts of data that are stored in databases or other forms of repositories. These repositories include multiple IT tools such as database technology, machine learning tools, information retrieval, and data analysis. Data mining can be characterised as the technology which incorporates the statistical techniques and mathematical equations that are used in an attempt to identify the significant relationships between variables in the historical data, to forecast or perform analysis on the data; or determine any significant relationship within the data captured[13].

Besides implementing new information technologies, libraries have to gain a better understanding of their user needs and expectations to serve them better[14]. To fulfil their users' requests and needs proactively, most libraries are performing their own research and analysis to get information to improve their services. One way of getting information of their users is by analysing the data captured in their database systems. This analysis is useful because it provides the library with information on what resources are being used, the quantity of resources and the demographics of the patrons. Although this research or analysis produces useful information, it is difficult to analyse the relationship of these data. The librarians who perform these analyses may not be experienced in performing a complete analysis or may just base their analysis on their experiences, which may not result in a complete and comprehensive analysis[7]. Data mining tool may be used to overcome the issues of performing analysis on any form of data. Data mining is the process of exploring and analysing large quantities of data to get meaningful information and make discoveries by the patterns generated in these data. The data mining processes consist of[7]:

(a) Defining the object of the research

(b) Identifying the data to be analysed

(c) Extracting and 'cleaning' the data

(d) Using the appropriate analysis tools

(e) Making discoveries from the patterns that arise from this process

(f) Analysing these patterns and generating reports.

Data mining has two main types of functions, which are descriptive mining and predictive mining[15]:

(a) Descriptive mining or discovery mining–the data mining process where useful information is extracted from the databases or, the discovery of potential knowledge from the databases.

(b) Predictive mining–data mining process that involves using machine learning to predict the unknown, or to obtain the future values of certain variables of interest from the data in the database or repositories.

The power of data mining is the ability of leveraging on the existing data of an organisation and using the information to make better strategic decisions[16]. The proven capability of the data mining system to produce information that is useful is the main reason why most organisations are implementing this system[17]. The data mining system utilises the data collection

in an organisation's data repoitories viz, legacy systems, databases, files, etc., which are in different format and structure. Therefore, it is important that the data is formatted and structured to suit the data mining activities to produce significant results[18]. The process of preparing the data for mining is called extraction, transformation and loading (ETL).

5. EXTRACTION, TRANSFORMATION AND LOADING (ETL)

The ETL is a data integration function that involves extracting data from a number of sources, transforming it to fit business needs, and ultimately loading it into a data warehouse for data mining[19]. An ETL process is the process of moving data from multiple sources, cleansing data, transforming and reformatting data, and finally loading the data to another database, data mart or data warehouse for data analysis[20].

6. BENEFITS OF DATA MINING FOR LIBRARIES

Currently, most libraries are performing analysis on the data captured in their database[21]. This analysis is useful because it provides the library with information on what kinds of resources are being used in the library or how these resources are being used. These types of analysis are driven by the librarians who perform the analysis based on their hypothesis or from their experience to get the statistical information. With the advancement of IT, a new data analysis process has evolved known as data mining. Data mining could solve some of the issues faced by the library by making knowledge discoveries from the library's data.

Thus data mining is the process of knowledge discovery from data, by producing patterns and trends of the raw data that may not have been analysed[10]. The libraries today are challenged to provide their patrons with the right source of information and services they require. Data mining can benefit the libraries by[7]:

(a) Improving the library services

(b) Being a support tool for decision-making by the library, and

(c) Producing reports which can help to justify decisions made.

6.1 Improve Library Services

Data mining can be used to improve the library services in the following[7] ways.

6.1.1 Collection and Borrowing Activities

Trends and patterns of borrowing of resources and materials could be provided by the data mining system. For example, the type of materials being borrowed in relation to the demographics of the clients would enable the libraries' to provide or increase the popular materials or resources for their clients.

6.1.2 Clients Information

Data mining can be performed on the clients' information or demographics which can produce patterns that may determine demographics of active clients, relationships between the clients and the materials demanded, behaviour pattern of the clients. This information is useful for the libraries to improve their services by providing the required services to the right groups of clients.

6.1.3 Web Logs

Data mining on the webpage log files could provide patterns on what, how and when the library sites are being used. This could provide information to the librarians on how to improve their websites.

6.1.4 Circulation Information

When data mining is performed on the circulation patterns of the library collection, it may provide the librarians with the knowledge of the future needs of their clients. This can guide the librarians on their planning, budgeting, purchasing of materials which are frequently used by their clients and can provide a justification for removing or archiving unused materials.

6.2 Decision-making Tool for Library Management

Data mining could help the libraries with decision-making in a number of areas, such as[7]:

6.2.1 Removal of Unused Material

By performing data mining on the pattern of circulation, materials that are not in circulation could be replaced with newer edition or the materials which are not in circulation could be stored somewhere else.

6.2.2 Staff Management

By performing data mining on the circulation of materials a pattern could be discovered on when and what type of services are in demand at a particular time. For example, if there is a pattern that shows there could be less renewal and borrowing of books in the morning, the librarian could be assigned to other library activities. This would ensure that the library staff are being fully utilised.

6.2.3 Purchasing of Materials

Data mining on the circulation could provide patterns on what types of books or materials are in demand and this may help the library in purchasing the 'right' materials for their clients. This information can be used to justify the purchasing of materials to the libraries' management.

6.3 Reporting and Justification

Library management, as most entities, have to provide reports and justification on decisions made. Data mining could provide the required information by[7] following reports:

6.3.1 Library's Usage

Data mining could provide management with projections of the future trends of the library usage. This could justify the need to increase the budget or resources allocated to the library.

6.3.2 Library's Staff Justification

Data mining can produce projection on the future requirement of the library and the growth of the library. The library management can justify and increase its work force based on this report.

7. CONCLUSIONS

With the data mining systems, libraries can have the right resources, better services, the ability to strengthen their collections and information for the decision-making process. The extraction, transformation and loading (ETL), a very important component in data mining, feeds quality data for data mining. The objective, strategy and budget of the data mining project should be the main criteria for selecting a suitable ETL tool to be used for extraction, transformation and loading activities. Exploration and analysis by automatic or semi-automatic means of large quantities of data in order to discover meaningful or relevant patterns and rules are of high advantage, competitive and time saving for libraries through the process of data mining. The libraries today are challenged to provide their clients with the right source of information and services they require, and this has been made possible with the technique and process of data mining.

REFERENCES

1. Conkling; *et al*. Engineering libraries: Building collections and delivering services. *Sci. Technol. Lib.*, 2001, **19**(3), 208.
2. Khosrowpour, Mehdi. Challenges of information technology management in the 21st century. *In* Information Resources Management Association International Conference, 2000. Idea Group Publishing, 2000.
3. Jiban, K. Pal. Usefulness and application of data mining in extracting information from different perspectives. *Annals Lib. Inf. Stud.*, 2011, **58**, 7-16.
4. Su, Di.; *et al*. Evolution in reference and information services: The impact of the internet. *The Reference Librarian*, 2001, **35**(74), 2.
5. Sun Microsystems Inc. Information technology advances in libraries. http://www.sun.com/products-n-solutions/edu/whitepapers/pdf/it_advances.pdf. (Last updated: 2003).
6. Neufeld, M. Lynne & Cornog, Martha. Database history: From dinosaurs to compact discs. *J. Amer. Soc. Inf. Sci.*, 1986, **37**(4) 183-190. doi: 10.1002/(SICI)1097-4571(198607)37:4<183::AID-ASI2>3.0.CO;2-W.
7. Nicholson, S. & Stanton, J. Gaining strategic advantage through bibliomining: Data mining for management decisions in corporate, special, digital, and traditional libraries. *In* Organisational Data Mining: Leveraging Enterprise Data Resources for Optimal Performance, edited by H. Nemati & C. Barko. Idea Group Publishing, Hershey, PA.
8. Okerson, Ann. Text and data mining: A librarian overview, IFLA WLIC, Singapore, Aug. 2013.
9. Ponniah, Raja. Date warehousing fundamentals: A comprehensive guide for IT professionals. John Wiley & Sons, Inc., New York, 2001.
10. The Free Dictionary. Farlex. http://www.thefreedictionary.com/?data+mining. (accessed on 4 June 2014).
11. Thomas, Connolly & Carolyn, Begg. Database systems: A practical approach to design, implementation, and management, Ed. 3. Addison Wesley, London, 2002.
12. Jiawei, Han & Micheline, Kamber. Data mining: Concepts and techniques, Ed. 3. Morgan Kaufman Publishers, San Francisco, 2001, p. 5, 40.
13. Becker, Shirley. Data warehousing and web engineering. Idea Group Publishing, Hershey, 2002.
14. Buckland, M. Five grand challenges for library research. *Library Trends*, 2003, **51**(4).
15. Zhengxin, Chen. Data mining and uncertain reasoning: An integrated approach. John Wiley & Sons, Inc., New York, 2001.
16. Nazem, S. & Shin, B. Data mining: New arsenal for strategic decision making. *J. Database Manag.*, 1999, **10**(1), 39-42.
17. Mulvenna, M.D.; Anand, S.S. & Buchner, A.G. Personalisation on the net using web mining. *Communications of the ACM*, 2000, **43**(8**)**, 122-25.
18. Charran, Eric. Introduction to data mining with SQL server, 2006. http://www.sql-server-performance.com/ec_data_mining.asp. (Last updated: 2006).
19. The Tech Dictionary. The concise Tech Dictionary, 2014. http://www.thetechdictionary.com/?page_id=195. (accessed on 5 June 2014).
20. Bitpipe. IT Information: White papers, product literature, webcasts and case studies (1998-2006), 2014. http://www.bitpipe.com/tlist/ETL-(Software).html/.

About the Authors

Mr Tariq Ahmad Lone is working as an Assistant Professor at The Business School, University of Kashmir, Srinagar. He is working in the area of Information Technology and is presently pursuing his PhD from Mewar University in the area of Image Processing. He obtained BSc from University of Kashmir and MCA from University of Jammu. He has published many research papers in professional journals in the areas of datamining, decision support system, and image denoising.

Dr Rafi Ahmad Khan is working as Scientist at The Business School, University of Kashmir, Srinagar. He is working in the area of Information Technology. He obtained BSc from University of Kashmir and MCA from University of Jammu. He obtained his PhD from University of Kashmir in the area of Business Intelligence. He has more than 20 research publications in national and international journals in the areas of business intelligence, datamining, and decision support system.