ANALYSIS OF WEB MINING APPLICATIONS AND BENEFICIAL AREAS

S. YADAV[1], K. AHMAD[2] AND J. SHEKAR[3]

CSE/IT Dept. S.I.T.E., SVSU, Meerut-250002, India

[1]Seema_yadava@yahoo.com, [2]khaleelamna@yahoo.co.in, [3]jayant_shekhar@hotmail.com

***ABSTRACT*:** The main purpose of this paper is to study the process of Web mining techniques, features, applications (e-commerce and e-business) and its beneficial areas. Web mining has become more popular and widely used in various application areas (such as business intelligent system, e-commerce and e-business). The e-commerce and e-business techniques run more efficiently with the application of mining techniques such as data mining and text mining. Among all the mining techniques web mining is considered best.

ABSTRAK: Tujuan utama kertas ini adalah untuk mengkaji proses teknik perlombongan Web, ciri-ciri, aplikasi (e-dagang dan e-niaga) dan bidang-bidangnya yang bermanfaat. PerlombonganWeb telah menjadi lebih popular dan digunakan secara meluas di pelbagai lapangan yang berbeza (seperti sistem perniagaan pintar, e-dagang dan e-niaga). Teknik-teknik e-dagang atau e-perniagaan berjalan dengan lebih lancar apabila teknik-teknik seperti perlombongan teks dan perlombongan data digunakan. Di kalangan semua teknik perlombongan, perlombongan web adalah yang terbaik.

***KEY WORDS**: Data mining, Web mining-commerce, e-business, Pattern discovery.*

1. INTRODUCTION

Mining the data over the World Wide Web (in short Web) using various data mining techniques and tools are known as Web mining. This area is most popular among researchers of today. The Web may be defined as the universal body of virtual data collection available for accessing information. This paper presents Web mining concepts and applications associated with electronic commerce and electronic business. Web mining is the most important application of data mining and other information processing techniques for finding useful patterns of data.

Web mining automatically extracts information from Web documents or services and can be applied to semi-structured or unstructured data like free-form texts. It focuses on many data types such as contents of Web pages, user access information, hyperlinks between pages and a variety of the Web resources to achieve intrinsic properties between the data objects. This is attained through methods such as machine learning, inductive learning and statistical analysis; and is useful as a tool for discovery and extracting information.

2. RELATED WORK

A study by Chakraborti [17] focuses on the problem areas of data mining related to hypertext, as well as on the problem of learning, in which a different methodology such as supervised/unsupervised learning and their implementation on hypertext. Srivastava *et al.* [12] in giving an overview of the system called 'WebSIFT' describe three phases of web usage mining as i. Preprocessing, ii. Pattern discovery and iii. Pattern analysis. *Sankar et al.* [18] highlight the limitations of and differences between data mining tools and web mining technologies in addition to a discussion on the future directions of using fuzzy logic (FL), artificial neural networks (ANNs), genetic algorithms (GAs) and rough sets (RSs)

Abraham [19] proposes 'Intelligent Miner (i-miner)' with fuzzy clustering and fuzzy inference system algorithms for complex e-commerce applications known as Business Intelligence (BI). A study by Renata and Vajk [20] deals with 'Frequent Pattern Mining' for Web Logs and discusses three pattern mining approaches from web usage mining point of view as i. Page sets, ii. Page sequences and iii. Page graphs. Much later, in proposing improvements to the Business Intelligent System applications, Atanasova *et al.* [4] present the solution to some problems in the marketing subsystem through a proposed functional matrix for the application of data mining, text mining, and web mining. Mei and Cheng [9] suggest the process of web mining techniques and its application towards electronic commerce. This paper concludes the relationship between electronic and web data mining, and gives the use of web mining technology in electronic commerce. Recently Yu and Yang [7] in discussing the methods and processes of data mining and their application in the field of electronic commerce, offer the proper classification of web mining, thus differentiating it from data mining.

3. CLASSIFICATION OF WEB MINING

Web mining can be divided or classified into four categories namely: i. Web Content Mining (WCM), ii. Web Structure Mining (WSM), iii. Web Usage Mining (WUM), and iv. User Profiles. The details are as below.

3.1 Web Content Mining (WCM)

This type of web mining is called as such because it is the process of discovery and extraction of useful information from the web page content, documents and multimedia data such as video, image, text, as well as audio discovered from the web. This type of web mining is also called web text mining, because text content is more popularly researched. The technology used in this type of web mining is known as Natural Language Processing and Information Retrieval for Information Gathering.

3.2 Web Structure Mining (WSM)

This is related to analyzing hyperlinks and link structures on the web for information retrieval and knowledge discovery of any type of data or items. Web structure mining can be used by search engines such as Google search engine; and is based on Page Rank

algorithm in which the relevance of a page increases with the number of hyperlinks to it from other pages. It is the process of finding useful knowledge using graph theory to analyze the connection and node structure of any web site. Web structure mining can be divided into two kinds (Fig. 1):

- Extracting patterns from hyperlinks in the web: A hyperlink is a structural component that connects the web page to a different location.
- Mining the document structure: Analysis of the tree-like structure of page structures to describe HTML or XML tag.

The technique which is used in this type of mining is HITS algorithm and the Page Rank algorithm.

3.3 Web Usage Mining (WUM)

This type of web mining is also referred to as web log mining, which is the process of extracting information from user data history from the web, either textual, or multimedia. It has become a necessary task for providing web administrators with meaningful information about users and usage patterns for improving quality of web information and service performance. Classification of WUM is given in Figure 1.
WUM applications include:

- Improving site design
- Targeting potential customers for electronic commerce
- Enhancing the quality and delivery of Internet information services to the end user
- Improving Web server system performance
- Identifying potential prime advertisement locations
- Facilitating personalization/adaptive sites
- Fraud/intrusion detection
- Predicting user's actions (allowing prefetching)

3.4 User Profile

User profiling web log mining provides demographic information about users of the Web site. This includes registration data and customer profile information of every user.

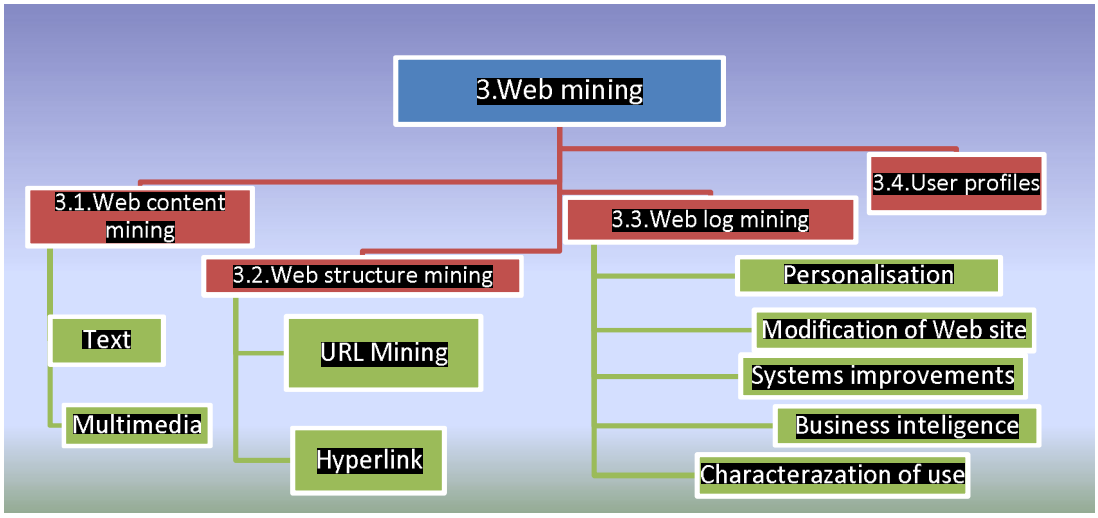

Fig. 1: Classification of web mining.

4. DIFFERENT PROCESSES IN WEB MINING

The processes of Web mining are divided into four stages: i. Source data collection, ii. Data preprocessing, iii. Pattern discovery, and iv. Pattern analysis.

4.1 Source Data Collection

The direct source of data in web mining is mostly web log files which are stored on the web server. Web log files record all the behaviour of the user which is connected at that time on the web, including the server log, agent log and client log.

4.2 Data Preprocessing

The data which generally gets collected from the web have features that are incomplete, redundant and ambiguous. Preprocessing provide accurate and concise data for data mining. This is the technique used to clean a server log files to eliminate irrelevant data, hence its importance for web log analysis in web mining. It includes data cleaning, user identification, user session certifications, access path supplements and transaction identification (Fig. 2), the details of which are as below.

4.2.1 Data Cleaning

This process removes the web log redundant data which is not associated with useful data thus improving the scope of data objects.

4.2.2 User Identification

The main work of this process is to identify the user uniquely on the web server. This can be done by cookies technology, user registration and heuristic rules.

4.2.3 User Session Identification

This process is done on the basis of user identification, and the purpose of this process is to divide each user access information into several separate sessions. The timeout estimation approach is used to separate each session process at the time web server is in use. In this way, when the allocated time is complete for one session, new sessions are initiated automatically.

4.2.4 Access Path Estimation

By the use of page caching technology and proxy servers, the access path is recorded by the access log files on the web server, which does not give the complete access path of users. Instead, path supplements can be archived by the use of web site topology to make the page analysis.

4.2.5 Transaction Identification

This process is totally based on the user session identification. Web transactions are divided or combined according to the demand of data mining tasks.

4.3 Pattern Discovery

There are many types of access pattern mining that can be used to perform on the need of the analyst. Some of these are given as path analysis, association rule discovery, sequential pattern discovery, clustering analysis and classification.

4.3.1 Path Analysis

The physical layout of any website is presented in graphical form. Web page is denoted as a node and the hyperlink between two pages is represented as edges in the graph.

4.3.2 Association Rules

The association rules are focused mainly in the discovery of relations between pages visited by the users on the website. Association rules can be used to relate the web page that is most often used by the single server session. These pages may be interlinked to one another by the help of hyperlinks. For instance, an association rule for a BBA program is BBA/seminar.html and BBA/speakers.html, whereby seminar is related to speakers.

4.3.3 Sequential Pattern Discovery

This technique is used to find inter-session patterns such that the presence of a set of items or data is followed by another item in an allotted time to that session. With the help of this approach, Web sellers or buyers can predict future visit patterns which will be helpful in placing advertisements aimed at certain user groups. There are also some other techniques which are useful on sequential patterns including change point detection, or similarity analysis and trend analysis.

4.3.4 Classification Analysis

Classification is the mapping of a data into one or several predefined data or items i.e. it classifies data according to the predefined categories. Classification can be done with the help of the following algorithms such as decision tree classifiers and naïve Bayesian classifiers. Web mining classification techniques allow one to develop a profile for clients who access a particular server files based on their access patterns.

4.3.5 Clustering Analysis

Cluster analysis technique is the most popular technique which is used in web mining, wherein a set of items or data which have similar attributes or characteristics is grouped together. It can help with marketing decisions of the marketers. Clustering of user information on the web transaction logs can develop a facility to future marketing strategies, both online and off-line. Two types of clustering methods are used namely: Hierarchical clustering (agglomerative vs. divisive and single link vs. complete link) and partition clustering (distance-based, model-based and density-based).

4.4. Pattern Analysis

Its main purpose is to find out a valuable model. Many types of techniques are used for analysis such as visualization tools, OLAP techniques, data and knowledge querying and usability analysis (Fig. 2), whose details are given below:

4.4.1. Visualization Techniques

This is a natural choice for understanding the behaviour of web users. The web is visualized as a directed graph with cycles represented as a node and hyperlinks denoted as an edge of the graph.

4.4.2. OLAP Online Analytical Processing Techniques

This is a very powerful paradigm for strategic analysis of database in business system. OLAP can be performed directly on top of the relational database.

4.4.3. Data and Knowledge Query

This is the most important analysis pattern in web mining; whereby focus is given on the proper analysis of user problems or user needs. Such type of focus is provided in two different ways:

- Constraints may be placed on the database in a declarative language.

- This query may be performed on the knowledge that has been discovered and extracted by the mining process.

4.4.4. Usability Analysis

In this analysis the details of software usability as well as user usability are given. This approach can also be used for any model for accessing behaviour of the user on the website.

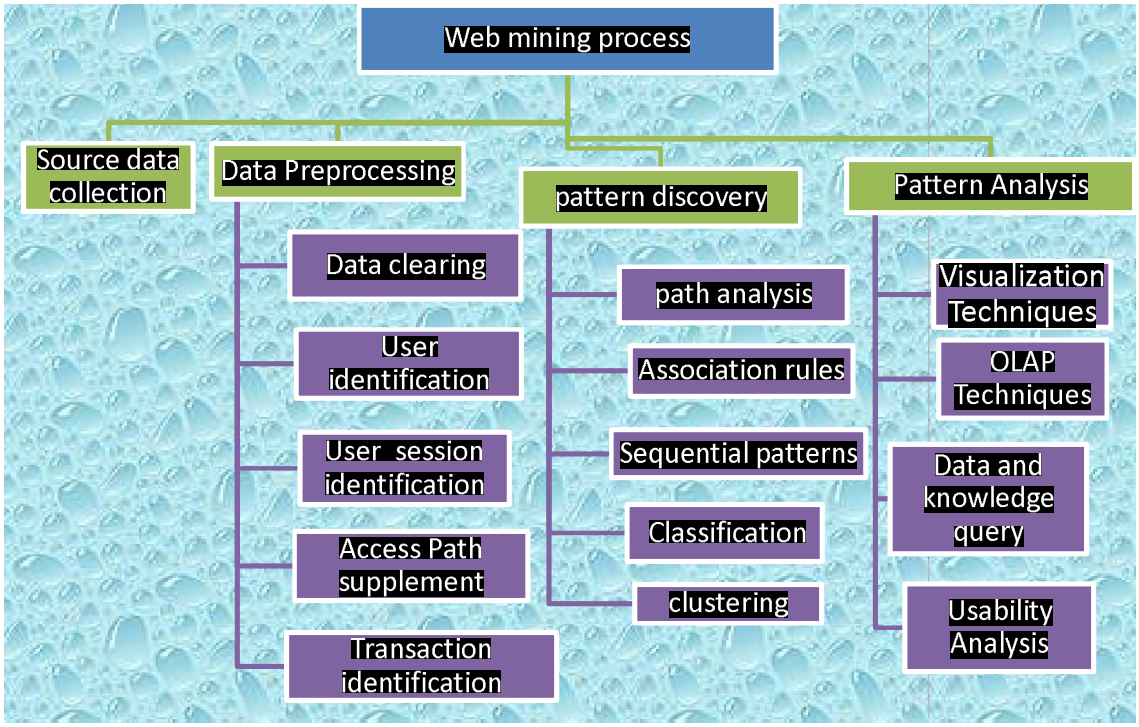

Fig. 2: Processes of web mining.

5. WEB MINING BENEFICIAL AREAS

There are several? benefit areas of web mining, some of which are given in Fig. 3 and briefly discussed below.

5.1 E-Learning

Web mining can be used for improving and enhancing the process of learning in e-learning environments. Applications of web mining to e-learning are usually web-usage based; i.e. online and not offline. Machine learning techniques and web usage mining enhance web-based learning environments.

5.2 Digital Libraries

Digital libraries services provide precious information distributed all around the world, eliminating the necessity to be physically present at different libraries in different parts of the world.

5.3 E-Government

Organizations that interact with citizens of the country lead to better social services. The main characteristic of e-government systems is related to the use of technology to deliver services electronically, focusing on the citizens' needs by providing better information and enhanced services in support of the government. E-government systems may provide customized services to citizens resulting in user satisfaction, quality of services and support in citizens' decision making, which lead to social benefits.

5.4 Electronic Commerce

A major challenge of e-commerce is to understand visitors' or customers' needs and to value orientations as such as possible. It can improve the capacity of the service for consumer and competitive advantages. The application of web mining in e-commerce is given in Fig. 3.

5.5 E-Politics and E-Democracy

E-politics provides political information and 'politics on demand' to the citizens improving political transparency and democracy. Election information, parties, members of parliament and members of local governments on the web are part of e-politics services. Despite the importance of e-politics in democracy there is limited web mining methods to meet citizens' needs.

5.6 Security and Crime Investigation

Web mining techniques are also used for protection of user system or logging information against such cyber-crimes as hacking, internet fraud, fraudulent websites, illegal online gambling, virus spreading, child pornography distribution and cyber terrorism. Clustering and classification techniques of web mining can reveal identities of cyber-criminals; whereas neural networks, decision tress, genetic algorithms and support vector machines can be used to trace crime patterns and network visualization on websites.

5.7 Electronic Business

Web mining techniques can support a web enabled electronic business to improve on marketing, customer support and sales operations. The applications of web mining in e-business are given in Fig. 3.

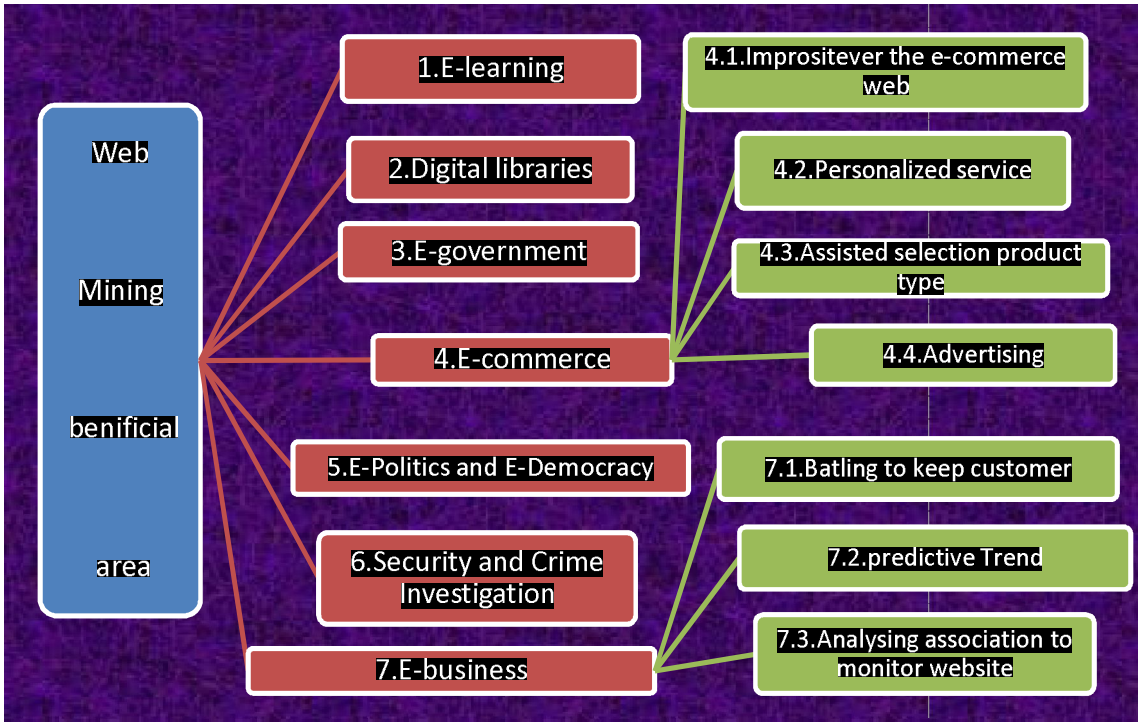

Fig. 3: Web mining beneficial areas.

6. CONCLUSION AND FUTURE WORK

Web mining methods have strong practical significance on e-systems. Web data mining forms the basis of marketing and e-commerce activities on the Web. It can also be used to provide fast and efficient services to users as well as building intelligent websites for businesses. Data mining in e-business will continue to be a very promising area of research.

Web mining enhances users' ability to access information hence the capacity and potentials of enterprise information resources can be fully reflected. It is expected that more applications of web mining will be developed.

REFERENCES

[1] Nacim Fateh Chikhi, Bernard Rothenburger, and Nathalie Aussenac Gilles, "A comparison of Dimensionality Reduction Techniques for Web Structure Mining" published in IEEE/WIC/ACM International Conference on Web Intelligence, 2007, pp. 116-119.

[2] Peter Owotoki, Nataša Manojlovic, Friedrich Mayer-Lindenberg, and Erik Pasche "A Data Mining Approach for Capacity Building of Stakeholders in Integrated Flood Management" Proceedings of the Sixth International Conference on Data Mining (ICDM'06) in IEEE, 2006.

[3] LIU Xiang-ying and YANG Lian-he "An Optimization Method of Network Teaching Management Based on Web Mining Techniques" published in 2nd International Conference on Future Computer and Communication in volume 3, 2010, pp. 348-352.

[4] T. Atanasova, M. Kasheva., S. Sulova and J. Vasilev, "Analysis of the possible application of Data Mining, Text Mining and Web Mining in Business Intelligent Systems" published in MIPRO, Opatija, Croatia, 2010, pp. 1294-1297.

[5] Li Haigang and Yin wanling, "Study of Application of Web Mining Techniques in E-Business" published in the IEEE conference in 2006, is supported by research project of NSFC (70571052).

[6] Du Ping and He YueShun, "Research of Remote sensing image Data Mining Technique based on Web" preceding in the IEEE computer society conference name "Asia-Pacific Conference on Information Processing", 2009, pp.298-290.

[7] HongKui Yu and Lijun Cao, Yuxiang Li, YanPing Yang, "Research Of Data Mining in Electronic Commerce" in IEEE Supported by Science and Technology Project of HeBei Science and Technology Department [Project Number: 07457253], 2011, pp.4323-4326.

[8] Alejandro A. Vaisman , Gabriel Dandretta and Mariela Sapia, "EnhancingWeb Access Using Data Mining Techniques" Proceedings of the 14th International Workshop on Database and Expert Systems Applications (DEXA'03), IEEE, 2003.

[9] Li Mei and Feng Cheng, "Overview of WEB Mining Technology and Its Application in E-commerce" preceding in 2nd International Conference on Computer Engineering and Technology, Volume 7, IEEE, 2010, pp. 277-280.

[10] MohammadReza Keyvanpour, Hamed Hassanzadeh and Babak Mohammadizadeh Khoshroo, "Comparative Classification of Semantic Web Challenges and Data Mining Techniques" preceding in IEEE conference (International Conference on Web Information Systems and Mining, 2009, pp. 200-203.

[11] Hui-Ju Wu, I-Hsien Ting and Kai-Yu Wang, "Combining Social Network Analysis and Web Mining Techniques to Discover Interest Groups in the Blogspace" is proceeding in IEEE international conference name Fourth International Conference on Innovative Computing, Information and Control, 2009, pp.1180-1183.

[12] Jaideep Srivastava_ y, Robert Cooleyz, Mukund Deshpande, and Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data" published in ACM SIGKDD Explorations. Copyrightc 2000 ACM SIGKDD, Volume 1, Issue 2, Jan 2000, pp. 12-23.

[13] Dr. A. C. Mondal and Sourav Maitra, "A Study of Web Mining Research – Last few years and the Road Ahead" published in ICCS, Burdwan University, 2010.

[14] Ling Chuanfan, "Application of Data mining in Electronic Commerce" from Google.

[15] Raymond Kosala and Hendrik Blockeel, "Web Mining Research: A Survey" proceeding in SIGKDD Explorations. Copyright c2000 ACM SIGKDD, Volume 1, Issue 2, Jan 2000. pp. 1-15.

[16] Wangbin Hu, Junpeng Yuan and Yuantao Song, "The Research of a Web Mining Method in Research Areas" published in the Sixth Wuhan International Conference on E-Business, e-Business Track, pp.314-319.

[17] Data mining for hypertext: A tutorial survey, Soumen Chakraborti, 2000.

[18] Sankar K. Pal, Varun Talwar and Pabitra Mitra, "Web Mining in Soft Computing Framework:Relevance, State of the Art and Future Directions", 2002.

[19] Arijit Abraham , "Business Intelligence from Web Usage Mining", 2003.
[20] Renata Ivancsy, Istvan Vajk,, "Frequent Pattern Mining in Web Log Data", 2006.