EXPLORATION OF EFFICIENT METHODOLOGIES FOR THE IMPROVEMENT IN WEB MINING TECHNIQUES: A SURVEY

Arvind Kumar Sharma*

Dr. P.C. Gupta**

ABSTRACT

Web mining is the application of data mining techniques to extract knowledge from the Web. Web mining techniques may be useful to automatically discover and extract meaningful information from Web documents. It is one of the prominent research area whose results mainly depend upon the proper preparation of the data from the Web logs. The web log data can be collected from Web servers, Proxy servers and Web clients. In this paper we have shown how web mining is implemented to obtain useful information through Web. We have also tried to identify the research area in Web mining where further work can be continued.

Keywords: *Web Mining, Web Usage Mining, Web Content Mining, and Web Structure Mining.*

*****Research Scholar, Jaipur National University, Jaipur, Rajasthan–India

Professor & Head,** Deptt. of Computer Science & Engineering, Jaipur National University, Jaipur, Rajasthan–India

1. INTRODUCTION

The World Wide Web (WWW) is the most heterogeneous and dynamic repository available. It is very popular and interactive. It has become an important source of information and services. The web is huge, diverse, expandable, scalable and dynamic[1,3]. Extraction of interesting information from Web data has become more popular and as a result of that web mining has attracted lot of attention in recent time[1]. Web mining is an application of data mining to large web data repositories[2]. A website is a collection of related web (World Wide Web) pages containing images, videos or other digital assets. Every website is hosted by at least one web server [24]. A web server is a program that, using the client/server model and the World Wide Web's Hypertext Transfer Protocol (HTTP), serves the files that form web pages to web users [23]. Web is a collection of billion of documents. Due to these features of web, we are currently drowning in information and facing information overload. So it is urgent and important to provide users with tools for efficient and effective resource and knowledge discovery on the web. Data in the World Wide Web is organized in inter-connected documents. These documents can be text, audio, video, raw data, and even applications[4]. Recent research focuses on utilizes the web as knowledge base for decision making. However, there is little work that deals with unstructured and heterogeneous information on the Web[5]. As per the web sites' survey more than 160,000,000 web sites are having inter, intra linked Web pages[6]. The speed of increase of web information is rapid. The hidden knowledge discovery, patterns and trends of user access can be found from the way the web sites and web pages are accessed and it is useful from the business perspective giving future directions for decision making. The web information is categorized into two categories: deep web and shallow web. This is shown in a fig. 1 below:

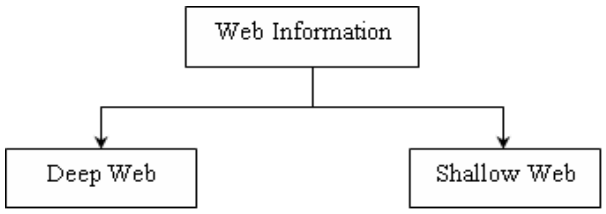

Fig. 1: Types of Web Information

The deep web includes information stored in searchable databases often inaccessible to search engines and it is accessed only by Web Site's interface. In other hand, the shallow web information can be accessed by search engines without accessing the web databases. Web mining is a natural combination of data mining and the WWW. It can be broadly defined as the discovery and analysis of useful information from the World Wide Web[6]. The information on the internet is in the form of static and dynamic web pages of various areas from education, industry to every walk of life including blogs. Web mining is a multi-disciplinary effort that draws techniques from fields like information retrieval, statistics, machine learning, natural language processing, and others[7].

Web mining is a special case of data mining. In section 2 categories of web mining is shown, section 3 discusses operations on web mining. Section 4 describes review of literature in which we have specified different work which has been done in web mining. Section 5 shows the methodologies which can be implemented for web mining. Section 6 concludes the paper while last section contains various references.

2. CATEGORIES OF WEB MINING

In this section we present taxonomy of web mining. The web mining is the use of data mining techniques to automatically discover and extract information from World Wide Web documents and services[13] in which at least one of structure or usage (web log) data is used in the mining process. Web mining can be divided into three distinct categories. This taxonomy is depicted in fig. 2.

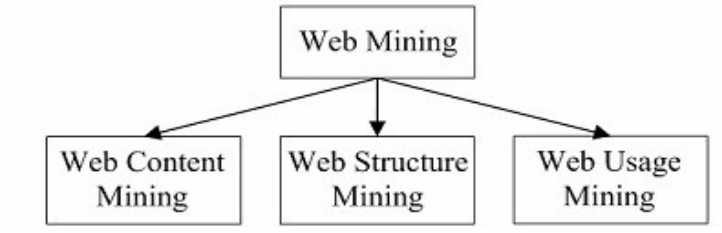

Fig. 2: Taxonomy of Web Mining

2.1 Web Content Mining

Web Content Mining is the process of picking up useful information from the contents of web documents. Content data is the collection of facts a web page is designed to contain. It

may contain text, images, audio, video or structured records such as lists and tables. Research activities in this field also involve using techniques from other disciplines[8] such as Information Retrieval(IR) and natural language processing(NLP).

2.2 Web Structure Mining

Web structure mining is a process of picking up information from linkages of web pages. It operates on the web's hyperlink structure. Web structure mining is also a process of using graph theory to analyse the node and connection structure of a web site. The structure of a typical web graph consists of web pages as nodes and hyperlinks as edges connecting between two related pages. In addition, the content within a web page can also be organized in a tree-structured format, based on the various Hyper Text Markup Language(HTML) and eXtensible Markup Language(XML) tags within the page.

2.3 Web Usage Mining

Web usage mining is also known as Web log mining. Web usage mining is the process of picking up information from user, how to use web sites. It is an application of data mining techniques to discover interesting usage patterns from web data, in order to understand and better serve the needs of web based applications. Usage data captures the identity or origin of web users along with their browsing behavior at a web site. Some of the typical usage data collected at a web site includes IP addresses, page references and access time of the users. The web usage data contains the data from Web server access logs, Proxy server logs, Browser logs, User profiles, Registration data, User sessions or transactions, Cookies, User queries, Bookmark data, Mouse clicks and Scrolls and any other data as the results of interactions.

3. OPERATIONS OF WEB MINING

- **Data or information extraction**: Our focus will be on extraction of structured data from Web pages, such as products and search results. Extracting such data or information allows one to provide services. Two main types of techniques, machine learning and automatic extraction are covered.
- **Web information integration and schema matching**: Although the Web contains a huge amount of data, each Web site (or even page) represents similar information

differently. How to identify or match semantically similar data is a very important problem with many practical applications. Some existing techniques and problems are examined.

- **Opinion extraction from online sources**: There are many online opinion sources, e.g., customer reviews of products, forums, blogs and chat rooms. Mining opinions (especially consumer opinions) is of great importance for marketing intelligence and product benchmarking. We will introduce a few tasks and techniques to mine such sources.
- **Knowledge synthesis**: Concept hierarchies or ontology are useful in many applications. However, generating them manually is very time consuming. A few existing methods that explores the information redundancy of the Web will be presented. The main application is to synthesize and organize the pieces of information on the Web to give the user a coherent picture of the topic domain.
- **Segmenting Web pages and detecting noise**: In many Web applications, one only wants the main content of the Web page without advertisements, navigation links, copyright notices. Automatically segmenting Web page to extract the main content of the pages is interesting problem. A number of interesting techniques have been proposed in the past few years [11].

4. REVIEW OF LITERATURE

Infinite web pages are either used or unused by users adding to large volume of space and their occurrence in web searches. 30-40% web pages are having duplication of the content approx. best estimate of unique static HTML pages is in billions from widely used search engines such as Yahoo, Google and increase continually. The following table 1 shows the facts of web sites increase from Nov 1995 to April 2011.

Table 1: Increase in the web sites from November 1995 to April 2011

Sr. No.	Web Site Survey Month & Year	Number of total web sites across all domains	Observations
1.	Nov 1995 to May	**Growth in Hostname and Active Web Sites**	Very high growth

	2000	There was an increase from 0 to 16,000,000. The growth was observed after year 2000.	
2.	May 2000 to May 2005	**Growth in Hostname and Active Web Sites** There was an increase from 16,000,000 to 64,000,000. The growth was observed from 0 to 32,000,000.	Rapid Increase in web sites. Active web sites indicating the presence of inactive web sites or web pages too.
3.	May 2005 to Feb 2008	**Growth in Hostname and Active Web Sites** There was an increase from 64,000,000 to 160,000,000. The growth was observed from 32,000,000 to approx. 70,000,000.	More increase in web sites and a high growth in host names.
4.	Feb 2008 till April 2011	**Growth in Hostname and Active Web Sites** The growth has been observed from 70,000,000 to approx. 312,693,296**.**	Further more increase in web sites requiring study of user access and behavior, link analysis of hyperlinks accessed by user adding value to business decisions.

Ref.: http://news.netcraft.com/archives/2011/04/06/april-2011-web-server-survey.html

In the last fifteen years, it has observed that the growth in number of Web sites and Visitors to those Web sites has increased exponentially. The number of Users by March 31, 2011was found 2,095,006,005which is 30.2% of the World's Population [21]. The number of Active Web sites is 130,493,668 as on June 13, 2011[22]. Due to this growth a huge quantity of web data has been generated.

Some more work has been recorded in web data mining till June 2011. The brief overview of the work is explained as below:

(1) Work on Web traffic mining through neural network was done since the task of mining useful information becomes more challenging when the Web traffic volume is enormous and keeps on growing. In this work, an intelligent model was proposed to discover and analyze useful knowledge from the available Web log data[14].

(2) Semantic web mining technique can provide meaningful search of data resources by eliminating useless information with mining process. In this technique web servers will maintain Meta information of each and every data resources available in that particular web server. This will help the search engine to retrieve information that is relevant to user given input string. This work combines the idea of these two techniques Semantic web mining and Probabilistic analysis for efficient and accurate search results of web mining[9].

(3) Most of the works were done using web mining and an online recommender system was proposed which was able to update incrementally and automatically the knowledge base obtained from historical usage data and to generate a list of links to pages (suggestions) of potentially interest for the users[15].

(4) Further more, the research work done and a model was proposed that focuses on an explorative study about data mining techniques suitable for understanding web usage patterns. Approach based on user profiles were found to be unreliable hard to understand dynamic surfer interests. Web page navigation pattern mining approaches were shown to improve performance of websites[16].

(5) Jyoti Pandey, et al [17] proposed data mining based service would run in background mode. The service computes the web pages likely to be requested by the user, considering their past web access log history, using association rules and thus optimizing the access time.

(6) In July 2007, I-Hsien Ting, Chris Kimble and Daniel Kudenko proposed a users "browsing behavior analysis approach which is based on applying web usage mining techniques. Two web usage mining techniques in the approach are introduced, including Automatic Pattern Discovery (APD) and Co-occurrence Pattern Mining with Distance Measurement (CPMDM). A combination method is also discussed to show how potential browsing problems can be identified [12].

(7) In the year 2010, a lot of works have done and a model was proposed to classify and match an online user based on his browsing interests using statistical techniques. A novel approach for recommendations of unvisited pages has been suggested in this

work. An offline data preprocessing and clustering approach is used to determine groups of users with similar browsing patterns.[18]

(8) In Jan 2011, the work done to provide an up-to-date survey of the rapidly growing area of web usage mining and how the various pattern discovery techniques help in developing business plans especially in the area of e-business. With the growth of Web-based applications, specifically electronic commerce, there is significant interest in analyzing Web usage data to better understand Web usage, and apply the knowledge to better serve users[5].

(9) In March 2011, some of the research works done and a model was proposed to observe that the various web data mining techniques which can support education system via generating strategic information. Since the application of data mining brings a lot of advantages in higher learning institution, it is recommended to apply these techniques in the areas like optimization of resources, prediction of detainment of faculties in the university, to find the gap between the number of candidates applied for the post, number of applicants responded, number of applicants appeared, selected and finally joined[19].

(10) In June 2011, a few works done and a model was proposed to prevent on data loss in website navigability. Data loss can be occurred due to noise and Remote File Inclusion (RFI) attack. By web log file it identifies the user activities on the website. In this research work, the proposed system used for securing website navigability through web mining[20].

5. METHODOLOGY

To facilitate web mining, there are a lot of techniques of web mining which can be applied to find patterns and trends in the data collected from the web. Few techniques are listed below:

5.1 Association Rules Mining(ARM)

Association rules mining is used to search correlation relationships among a large set of data items or variables. The association rules can be seen as the identification of actions or facts that, being initially independent, they happen in a combined or associate way. The considered facts can be characteristics or behaviors observed in the individuals. A typical example of association rules mining is the market basket analysis[7] i.e. a supermarket might gather data

on customer purchasing habits. Using association rule learning, the supermarket can evaluate which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as Market Basket Analysis.

5.2 Classification

The classification is the task of generalizing known structure to apply to new data. It is also used to predict the classes of future object or data. In web mining, classification rules allow one to develop a profile of items belonging to a particular group according to their common attributes[10]. For example, an e-mail program might attempt to classify an e-mail as legitimate or spam. It is a two-step process. In the first step, a classification model is built based on training data set. In the second step, the model is applied to new data for classification.

5.3 Clustering

Clustering is a technique to group together a set of items having similar characteristics. It is a unsupervised learning process. It is the process of grouping a set of physical or abstract objects into classes of similar objects, so that objects within the same cluster must be similar to some extend, also they should be dissimilar to those objects in other clusters. In clustering, objects are grouped together based on their similarities [9] and the aim of clustering is to assemble vast data items into classes to acquire least similarity degree between classes and maximal similarity degree in a class.

6. CONCLUSION

The Web (WWW) is huge, diverse and dynamic. It is a collection of billion of documents. It is urgent and important to provide users with tools for efficient and effective resource and knowledge discovery on the web. Recent research focuses to utilize the web as knowledge base for decision making. However, there is a little work that deals with unstructured and heterogeneous information on the Web. Web mining is the application of data mining techniques to extract knowledge from Web data, where at least one of the structures (hyperlink) or usage (web log) data is used in the mining process. In order to improve the efficient methodology offered by the dataset, utility of the data suffers. On conducting the

experiments we will find an efficient methodology for the improvement in web mining techniques.

7. REFERENCES

[1] Cooley, R. and et al "Web mining: information and pattern discovery on the World Wide Web" International Conference on Tools with Artificial Intelligence, Newport Beach, IEEE, 1997, pp. 558-567

[2] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava "Data preparation for mining World Wide Web browsing patterns" Journal of Knowledge and Information System 1999, pp. 1-27

[3] Navin Kumar Tyagi & et al. "Analysis of Server Log by Web Usage Mining for Website Improvement" International Journal of Computer Science Issues, Vol.7, No. 8, July 2010

[4] Osmar R. Zaïane "Principles of Knowledge Discovery in Databases" CMPUT690, pp.9 1999

[5] Sachin Yele & et al "Web Usage Mining for Pattern Discovery" International Journal of Advanced Engineering & Application, page 19-20, Jan 2011

[6] Pradnya Purandare "Web Mining: A Key to Improve Business on Web" IADIS European Conference Data Mining, page 155-157, 2008

[7] Srivastava, J., Cooley, R., Deshpande, M., Tan, P. "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data" SIGKDD Explorations, 2000 Paper available on: http://www.acm.org/sigs/sigkdd/explorations/issue1-2/srivastava.pdf

[8] S. K. Pani "Web Usage Mining: A Survey on Pattern Extraction from Web Logs" International Journal of Instrumentation, Control & Automation (IJICA), Vol.1, Issue 1, 2011

[9] T. Krishna Kishore, T. Sasi Vardhan & N. Lakshmi Narayana "Probabilistic Semantic Web Mining Using Artificial Neural Analysis" International Journal of Computer Science and Information Security, Vol.7, No.3, March 2010

[10] Sheilini jindal et al; Research support systems as an effective web based information system" Journal of Global Research in Computer Science, 2 (5), may 2011

[11] S. Mukthyarazam, M. Kiran Kumar, Shaik Rasool & S. Jakir Ajam "Web data mining Using XML and Agent Framework" International Journal of Computer Science and Network Security, Vol.10 No.5, May 2010.

[12] Natheer Khasawneh, Chien-Chung Chan "Active User-Based and Ontogy-Based Web Log Data Preprocessing for Web Usage Mining" Dec 2006

[13] Jaideep Srivastava et al. "Web usage mining: Discovery and applications of usage patterns from web data" SIGKDD Explorations, 1(2):12–23, 2000

[14] Farhad F. Yusifov "Web Traffic Mining using Neural Networks" World Academy of Science, Engineering and Technology, 21 2006

[15] Rajni Pamnani, Pramila Chawan "Web Usage Mining: A research area in Web mining" International Conference on Recent Trends in Computer Engineering, ISCET 2010, RIMT, Punjab, ISBN 978-81-910304-0-2

[16] Jin Xu, Yingping Huang and Gregory Madey; University of Notre Dame, IN 46556; A research support framework for web data mining; Oct 13, 2003, Halifax.

[17] Jyoti Pandey, Amit Goel, Dr. A K Sharma "A Framework for Predictive Web Prefetching at the Proxy Level using Data Mining" IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.6, June 2008

[18] C.P. Sumathi "Automatic Recommendation of Web Pages in Web Usage Mining" International Journal on Computer Science and Engineering, Vol.02, No. 09, 2010

[19] Dr. Varun Kumar, Anupama Chadha "An Empirical Study of the Applications of Data Mining Techniques in Higher Education" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No.3, March 2011

[20] Kavita Sharma et al. "Private and Secure Hyperlink Navigability Assessment in Web Mining Information System" International Journal on Computer Science and Engineering (IJCSE), ISSN : 0975-3397 Vol. 3 No.6 June 2011

[21] http://www.internetworldstats.com

[22] http://www.domaintools.com/internet-statistics

[23] James B. Lingan, http://whatis.techtarget.com seen on March 2011

[24] http://en.wikipedia.org seen on March 2011.