Tools and Techniques of Web Mining – Survey

T. Pavithra[#1,] K. Thangadurai[#2]
[#1]Research Scholar, Government Arts College, Karur, Tamil Nadu, India.
E-mail: pavithrakrishnaveni2011@gmail.com
[#2]Asst. Professor and Head, PG and Research Department of Computer Science,
Government Arts College, Karur, Tamil Nadu, India.
E-mail: ktramprasad04@gmail.com

Abstract: *Web Mining is move towards the World Wide Web and more useful environment in which users can access the information quickly and easily which they need. There is huge amount of text documents, multimedia files and images are available in the web and it is still increasing. Data mining has involved an excellent deal of attention within the data business and in society such data into helpful information and knowledge. The information and knowledge extended to be used for applications starting from market research, fraud detection, and client retention, to production management and varied explorations. This paper is to examine the role of data mining for data extraction in web page, structure and usages mining which concerned with its techniques, tools and applications.*

Keywords: *Data Mining, Data Extraction, Web mining, Web content mining, Web structure mining, Web usage mining.*

I. Introduction

Web mining is the process of data mining techniques and algorithms to abstract data directly from the Web, by extracting it from Web documents and services, Web content, hyperlinks and server logs. (Mining means dig out something useful or valuable from a baser substance, such as mining gold from the earth). The goal of Web mining is to look for patterns in Web data by collecting and analysing information in order to gain insight into trends, the industry and users in general. Use of mining data will give most effective to those paths to the portals. Web mining is a branch of data mining concentrating on the World Wide Web as the main data source, including all of its components from Web content. The term Web mining has been used in three different ways. The first, Web content mining is the process of data discovery from sources through the World Wide Web. The second, Web structure mining is the process of analysing the relationship between Web pages which is linked by information or direct link connection through the use of graph theory. The third, Web usage mining is the process of extracting patterns and information from server to get addition on user activity.

II. Categories Of Web Mining

- **Web content mining** —This is the process of mining useful information from the contents of Web pages and Web documents, which are mostly text, images and audio/video files.

- **Web structure mining** —This is the process of analysing the nodes and construction of a website through the use of graph theory. There are two things that can be found from this: the structure of a website in terms of how it is connected to other sites and the document structure of the website by its own and each page is connected.

- **Web usage mining** — This is the process of extracting patterns and information from server logs to gain insight on user activity including where the users are from, how many click and what item on the site and the types of activities is done on the site.

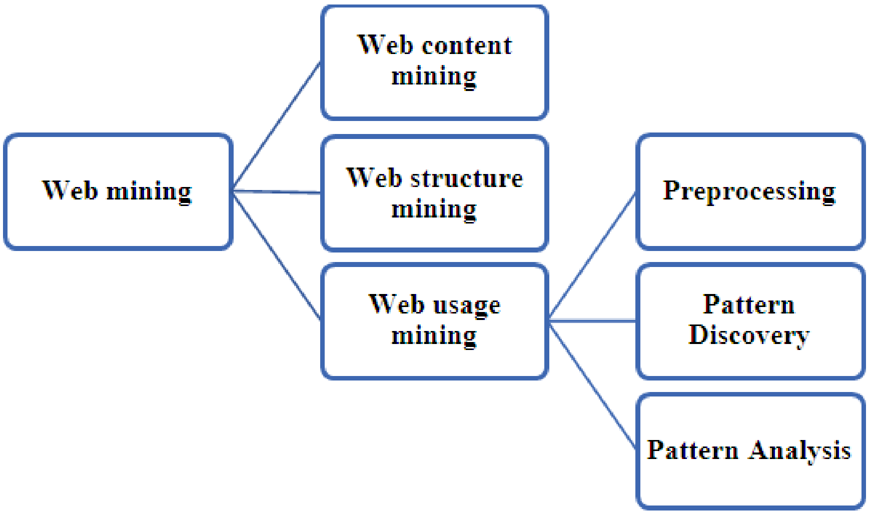

Figure 1: Web mining classification

III. Application of web data processing

Web data processing is such that fixed information is extracted and also the following options incorporated with the web mining program should be fastened if we want to use data processing with success in making web intelligence. Web mining is striking for enterprises because of several advantages. This not only helps companies to retain customers by being able to provide more modified services, but it also pays in the search for potential customers.

- **Web search-engine data processing**

There are three basic components they are as follows:

- Web Crawler.
- Database.
- Search Interfaces.

The repository of web pages is otherwise called as the 'Index', and this data store which is organized and used to provide the search results you see on the search engine. Indexing is the process of organizing the multitudes of data and pages so they can be searched speedily for related results to your search query.

- **Web Link Structure determine**

A website structure shows the website and its set up, i.e. the individual subpages are linked to any another. It is particularly important that creeps can find all subpages fast and easy when websites have a huge number of subpages.

- **Dynamic web Mining**

Mining Web accessibility activities which have the opposite hand, possible in several programs and helpful. Customers will mine the data to get website access. Assessing and determining regularities in a huge data will enhance the standard and distribution of web data and services to the top individual, enhance web hosting, server system performance and acknowledge customers for electronic trade.

- **Sequential Pattern Discovery**

Sequential pattern mining is concerned with finding statistically appropriate patterns between information examples where the data are delivered in order. It is usually reputed that the values are detached, and thus time series mining is closely related, but usually considered a different activity.

IV. Challenges in Web Mining

- **The web is too vast** −The size of the web is massive and fast increasing. This seems that the web is so huge for data warehousing and data mining.

- **Complexity of Web pages −**The web pages do not have merging structure. They are very difficult as compared to traditional text document. There is enormous amount of documents in digital library of web. These libraries are not arranged according to any exact sorted order.

- **Web is dynamic information source −**The data on the web is fast updated. The data such as weather, sports, stores, etc., are regularly updated.

- **Diversity of user communities −** The user community on the web is fast expanding. These users have different backgrounds, interests, and usage. There are more than million workstations that are connected to the Internet and still fast increasing.

- **Relevant Information −** It is considered that a particular person is generally interested in only small portion of the web, while the rest of the portion of the web contains the data that is not relevant to the user and may marsh desired results.

V. Web Mining Tools

Different types of tools are used to mining the data. Some of these tools are as follows.

- **WebIESoft**

It is a most powerful tool for web data mining, content extraction and content update monitor. It can extract structure or unstructured data from web page, which change to local file or save into database, post to web server. No need to define complex template rules, just browse to the web page you are interesting and click what you want to extract, and run it as you want, or let it run automatically.

- **Mozenda**

Mozenda is a technology which delivered as either software or as a managed service, that allows people to capture unstructured to a structured format. Web scraping is the automatic process of mining data or collecting information from the World Wide Web.

- **Web scraping**

Web scraping is a method to get data from a website. Web scraping tool is otherwise called as a website scraper, which is able to extract lots of data through an automatic process. The tool mechanism by sending a query to the requested pages, then combing through the HTML for specific items.

This is tool which combines predictable RPA with intellectual elements like natural language understanding and reading any unstructured data. This tool allows organizations to automate the processes which are performed by the humans.

Table 1: Some tools used in Web mining

TOOLS	FEATURES
Data Pre-Processing Tools	
DataPreparator	Performs cleaning, extraction and transformation of data before pattern discovery.
Sumatra TT	Platform independent data transformation tool. Based on Sumatra script and supportRapid application Development.
Pattern Discovery Tools	
i-Miner	Discover data cluster by using fuzzy clustering algorithm and fuzzy inference system for pattern discovery and analysis
Argunaut	Develop the patterns of useful data by using sequence of various rules.
Pattern Analysis Tools	
Webalizer	GNU GPL license based and produces web pages after analysing patterns.
Naviz	Visualization tool that combines 2-D graph of visitor access and grouping of related pages. It describes the pattern of user navigation on the web.
Stratdyn	Enhances WUM and provides visualization of patterns.

VI. Conclusion

In this paper we termed some of tools available to work on web mining prominent tools/techniques for the Web Content Mining, Web Structure Mining and Web Usage Mining. We analysed the strength and precincts to provide comparison among them. This approach generates quality endorsements by evaluating collective effort instead of basic recommendations on just one person's past expertise. Actually, collective filtering has been used as a knowledge mining technique for web data processing and effective result presentation in future.

References

[1]. R. Chau, C. Yeh and K. Smith, Personalized multilingual web content mining, KES (2004).

[2]. B. Liu and K. Chang, Editorial: Special issue on web content mining, SIGKDD Explorations 6(2) (2004).

[3]. Ricardo Baeza-Yates and Alessandro Tiberi. ―Extracting semantic relations from query logs proceeding for ACM SIGKDD international conference on Knowledge discovery and data mining, 2007.

[4]. P. Kolari and A. Joshi, Web mining: Research and practice, Comput. Sci. Eng.July/August (2004).

[5]. Bodyan G.C, Shestakov T.V, "Web Mining in Technology Management", Engineering Universe for Scientific Research and Management, Vol 1 Issue 2, April 2009.

[6]. Preeti Chopra, Md. Ataullah, a Survey on Improving the Efficiency of Different Web Structure Mining Algorithms in International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 –8958, Volume-2, Issue-3, February 2013.

[7]. Kosala and Blockeel, ―Web mining research: A survey, SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery and Data Mining, ACM, Vol. 2, 2000.

[8]. Screen-scraper, http://www.screen-scraper.com Viewed 19 February 2013.

[9]. Zhang, Q., Segall, R.S., Web Mining: A Survey of Current Research, Techniques and Software, International Journal of Information Technology & Decision Making. Vol.7, No. 4, pp. 683-720. World Scientific Publishing Company (2008).

[10]. Q. Yang and X. Wu, 10 challenging problems in data mining research, International Journal Information Technology Decision Making 5(4) (2006).

[11]. Andrei Broder. A taxonomy of web search. SIGIR Forum, 36(2):3–10, 2002.

[12]. Jiawei Han, Kevin, Chen-Chuan Chang "Data Mining for Web Intelligence" IEEE International Conference on Data Mining, 2002.

[13]. Qingyu Zhang and Richard s. Segall, Web mining: a survey of current research, Techniques, and software, in the International Journal of Information Technology & Decision Making Vol. 7, No. 4 (2008).

[14]. J. Srivastava, R. Cooley, M. Deshpande and P. Tan., "Web Usage Mining: Discovery and Applications of Usage Patterns from Web data", Department of Computer Science and Engineering, University of Minnesota. SIGKDD Explorations, 1(2):12, January 1999.

[15]. N. Barsagade, Web usage mining and pattern discovery: A survey paper, Computer Science and Engineering Dept., CSE Tech Report 8331 (Southern Methodist University,Dallas, Texas, USA, 2003).

[16]. J. Srivastava, R. Cooley, M. Deshpande, and P.N. Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations, vol. fI, no. 2, pp. 12-23, 2000.