Advances in Computer Science and Information Technology (ACSIT)
Print ISSN: 2393-9907; Online ISSN: 2393-9915; Volume 1, Number 2; November, 2014 pp. 56-60
© Krishi Sanskriti Publications
http://www.krishisanskriti.org/acsit.html

Web Mining: Taxonomy and Survey

Surabhi Singh[1], Devyani Gupta[2] and Aakash Chandhoke[3]

*[1]Department of Computer Science and Engineering Student at JSS Academy of Technical Education
C-20/1, Sector 62 Noida, Uttar Pradesh
[2]Department of Computer Science and Engineering Student at JSS Academy of Technical Education
C-20/1, Sector 62 Noida, Uttar Pradesh
[3]Department of Computer Science and Engineering Student at JSS Academy of Technical Education
C-20/1, Sector 62 Noida, Uttar Pradesh
E-mail: [1]ss_yadav7@yahoo.co.uk, [2]devyanigupta10@gmail.com, [3]aakash.chandhoke24@gmail.com*

Abstract: *Internet, the hub of voluminous and heterogeneous data, has become a vital part of our lives. With an immense amount of data available on the internet, web mining has emerged as a significant area of research due to its wide range applications. The term web mining was coined by Etzioni in the year of 1996. Web mining primarily concerns with application of data mining techniques to extract the patterns of data on the web and to gather useful information. The field of web mining can be classified into three distinct branches i.e. Web Content Mining (WCM), Web Usage Mining (WUM) and Web Structure Mining (WSM). Web Content Mining is the process of extracting useful information from the contents of Web documents. Web Structure Mining can be defined as the process of discovering structure information from the Web .Web Usage Mining is the discovery of meaningful patterns from data generated by client-server transactions on one or more Web localities. The aim of this paper is to provide an overview of Web Mining and its applications as well as an evaluation of its categories. Furthermore, we will also be outlining a promising future research in the field of Web mining i.e. Cloud Mining.*

Keywords: *Web Mining; Web Content Mining; Web Usage Mining; Web Structure Mining; Cloud Mining.*

1. INTRODUCTION

The internet has evolved into a global information space for almost all types of information that can be represented in various formats such as text, audio, video, graphics and animation. Internet is being used for hosting data and information pertaining to various application that are used in all dimension of humankind, such as, in aviation, transportation, military and space information, meteorological system, medical and health, agriculture and education etc. Due to phenomenal amount of data being kept on the World Wide Web, it has eventually become the prime source of information. Millions of users navigate through the World Wide Web each day to retrieve useful information from it. To retrieve the best results from the enormous amount of data available on the web, an application of data mining known as web mining is used. In traditional data mining, processing million records from a database requires good amount of time, however, using web mining techniques facilitate the processing of even higher amount of record by eventually requiring little time to process. Data mining requires retrieval of information from large structured databases, however, in web mining, the data and information is acquired from semi structured or unstructured web pages. Furthermore, when data mining of corporate information is done, the data is private and generally requires access rights to read. Whereas, the data used in web mining is public and rarely requires access right.

2. HISTORY

Web mining is defined as a process of extracting useful information from the web so that in future better web applications may be developed for the benefit of users. Web mining requires exhaustive research on developing data patterns for which the mining is carried out. This advancement requires exploration of how the data is stored on the internet and continues to evolve in data admittance and in real time manner. Data that are collected from various surveys or inputs from various network sources from computers and tapes are known as Data collection [1]. Data storage refers to the storage of information and efficient retrieval of that stored information. Web mining is an extended and advanced form of data collection. The main techniques used in web mining have been benefited due to an extensive development in the supporting areas such as, machine learning and artificial intelligence. As discussed in [1], chronological developments in the area of web mining are outlined in following table.

Mile Stones	Product Providers	Empowering Technologies
Data Storage on file based systems (1960s,70s)	IBM, CDC Main frame products	Tapes, Disks
Data Retrieval and Processing (1980s)	Oracle, Sybase, IBM, Microsoft, Informix	Relational Database (RDBMS), Structured Query language(SQL), ODBC

Data Warehousing, Decision Support systems (1990s)	Pilot, Arbor, Micro strategy	On-Line Analytic Processing, Multidimensional Database
Data Mining (2000s)	Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry)	Advanced algorithms, parallel processing Computers, Massive Databases
Web Mining (as on date)	IBM, Fuzutsu, Big Data, Cloud Computing, SPSS, Net Genesis	WWW, Internet, monumental scale Database

Fig. 1: History of Web Mining

3. CLASSIFICATION

Web mining can be classified [1,2] into three distinct areas namely Web Content Mining, Web Structure Mining and Web Usage Mining.

3.1 Web Content Mining

Web Content Mining (WCM) can be described as a process of extracting valuable information from the content of a web page. The web page consists of data of various types such as unstructured data like text, semi-structured data which includes HTML and more structured data like data in tables. Data Mining and Text Mining are the two terms which are related with WCM. Data mining is associated with WCM as data mining techniques can be applied to it. Moreover, text mining is related to it since majority portion of the web contents are text.

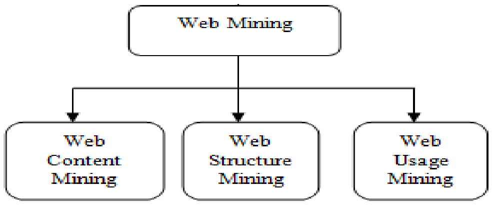

Fig. 2: Classification of Web Mining

Web Text Mining is extremely efficient whenever it is used in relation to a content database which deals with a precise topic. For instance, online universities employ a library system to summon up articles associated to their general areas of study. This particular content database allows us to retrieve only the information of those particular subjects. Hence, it provides the most precise outcome of search queries on search engines. Since it offers the most significant results, the quality of search results is very high. The two points of view from which Web Content Mining can be differentiated are the Information Retrieval View and the Database View.

Advancement in techniques for done for semi-structured data and unstructured data from the view point of Information Retrieval are outlined in [2]. It shows that to represent unstructured text, most of the researchers use bags of words. For the semi-structured data, all the researches make use of the HTML structures present inside the documents and some of them made use of the hyperlink structure between the documents to represent text documents. Whereas in database view, the mining tries to deduce the web site structure or transform it to a database in order to have better management of the information. This can be achieved by building a virtual database, building a web warehouse or building a web knowledge base. A number of techniques of data mining are used to mine the web content. One such technique is classification. Classification is a supervised technique [3], in which a number of categories are defined and the web documents are allocated to one or more categories. In the “training” period, a term vector is created and the “definition” of a category is generally in the form of such term vector. Another technique is clustering which is an unsupervised technique. In this technique, documents which are similar are grouped to aid the user to choose a topic of interest. Summarization is also a technique of web content mining in which we reduce the length of the document to aid the user to make their mind up whether the document is to be read or not.[4] Furthermore, there is another technique known as topic tracking. In this technique, the interest of the user is tracked. The documents which have been viewed by the user are checked and attempts are made to locate other documents which may be related. This technique is generally used by the registered sites. An example of a site using topic tracking technique is Yahoo.

The data available on the web is voluminous as well as heterogeneous. Some data are more relevant as compared to other data whereas some are less relevant. There are four criteria on the basis of which the relevance of the content is measured- document, query based, user based, and role/task based. The criterion of document for relevance is generally used in context of queries wherein the results are prepared on the foundation of some relevance. The most common measure of relevance is query based relevance. It is conventional in Information Retrieval. User based relevance is generally allied with personalization. For a particular user, a profile is developed and the correspondence between a profile and document is calculated. The fourth measure of relevance which is role/task based is similar to user based relevance. The only difference between the two is that in User based relevance the profile is based on an individual whereas it is based on a particular role in case of role/task based.

3.2. Web Structure Mining

Web Structure Mining (WSM) is defined as the process of generating the structure of links of the web pages. It discovers the hyperlink structure in inter documents level of the web page and finds the similarity and relationship among different web pages. It is of great importance to understand web data structure for retrieval of information [3]. The aim of Web Structure Mining is to generate abstract of the website and web pages. In this field the number of out links, i.e. links from web, and the number of in links (in edges), i.e. links towards the web pages are of utmost importance. The attractiveness of the web page is generally measured by the fact that a particular web page should be referred to by a large number of other web pages and the significance of web pages may be adjudged by a large number of out links contained by a page. The more links provided within the association of a web page enables the navigation to produce link hierarchy allowing easy navigation [9]. Therefore in the field of web mining, web structure mining has become a very imperative area for further study.

Conventionally the web documents contain links and they use the primary data on the page. So, it can be said that Web Structure Mining has a relation with Web Content Mining. These two mining techniques are being widely used together in applications. Web structure mining is essential because of the provision of the web structure schema through database techniques for web pages. It extracts previously unknown relationships between web pages. It allows data to be drawn from the search engine and links the web page from the web site where the contents are kept, using a search query. The completion takes place with the help of spiders scanning the web sites and linking the information through reference links to show the web page with the desired material. Web mining minimizes two main problems of the web. The first problem is inappropriate search result. Relevance of information searched becomes unorganized. This is due to the tolerable capacity of search engines. The second problem is the inability to index large amount of information on the web. This causes low amount of recall with content mining [8].

Therefore web mining and its branch structure mining can provide tactical results for marketing of a web site for sale of production. The more traffic directed to the site increases the visitation to that site again and again, thus increasing the traffic and increasing database entry for that site which is valuable for it. This enables the marketing strategies to provide solutions to the problem that are more productive with decreasing efforts. Using this concept, various web pages can be ranked based on the search query or the relevance of data that it contains. Thus, it decreases the efforts of surfing through multiple pages and saves the user from reading information that is not valuable to them.

3.2.1. Webpage Ranking. Surfing the internet involves two main practices. Extracting the information from the network and ranking them according to their eminence. There are many gadgets and tools used for efficient searching. Thus page ranking is an important field for information retrieval. Search engines now-a-days return millions of web pages for a single query [6]. It is not possible for the users to surf through all the web pages and links in that web pages.

Rankers are classified into two groups: Connectivity-based rankers and Content-based rankers [7]. Connectivity-based rankers work on the basis of link analysis technique; links are edges that point to different web pages. Content-based rankers work on the basis of number of matched terms, location of terms, frequency of terms, etc. Page Rank provides more efficient way to calculate the rank of the web pages. Page ranking is not decided by merely counting the number of back links. If a backlink comes from an important page then that hyperlink is given more weightage than the backlink from non-important pages [7].

Many algorithms have been proposed for page ranking. Important among them is Page Rank algorithm and Hypertext Induced Topic Selection (called as *"HITS"*). Page Rank algorithm has been developed by Google and is named after its co-founder and president Larry Page [7]. It uses the universal link information and is stated to be the primary link approval scheme employed by the most favorite *"Google Search*". HITS algorithm is based on the hyperlink developed by John Kleinberg [6]. It decides the ranking based on the textual contents against the given search query. It concentrates on structure of the web only, neglecting the content, and ranks them by in links and out links. In this algorithm, Authorities are the web pages pointed by many hyperlinks and Hubs are the pages that point to many hyperlinks [5].

3.3. Web Usage Mining

Web Usage Mining (WUM) is the discovery of patterns in clickstreams and associated data collected or generated when users access the websites [10]. In recent years, there has been a flourishing increase in number of researches on Web Mining and specifically on Web Usage Mining. Since the mid-1990s, more than 400 papers have been published on Web Mining, out of which about 150 papers were published before 2000; out of these papers around 50% papers regarded Web Usage Mining.

Web Usage Mining is categorized into three stages (i) preprocessing, (ii) pattern discovery and (iii) pattern analysis.

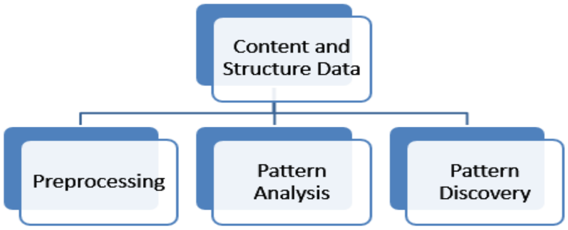

Fig. 3: Steps of Web Usage Mining

Pattern discovery is representing methods and algorithms of various different realms like statistics, data mining, machine learning and pattern recognition in the form of rules, tables, charts, graphs, or any other visual presentation forms for characterizing, categorizing or relating data from the web access logs [11].

Pattern analysis is filtering of unnecessary patterns found in the pattern discovery stage. The most commonly used pattern analysis is in the form of SQL queries [11].

Data preprocessing has a significant role in Web Usage Mining applications. It comprises four different methods (i) Data Cleaning, (ii) Session Identification, (iii) Retrieval of Information, and (iv) Data Formatting. Data Cleaning is the removal of all the irrelevant data monitored in web log, e.g.: requests for graphical page content. Session Identification (also known as sessionization) identifies the activities of users' session from the moment user enters the site until it leaves it.

Content and Structure Retrieving improvises the information extracted from web access logs. Data Formatting is the proper formatting of data [13].

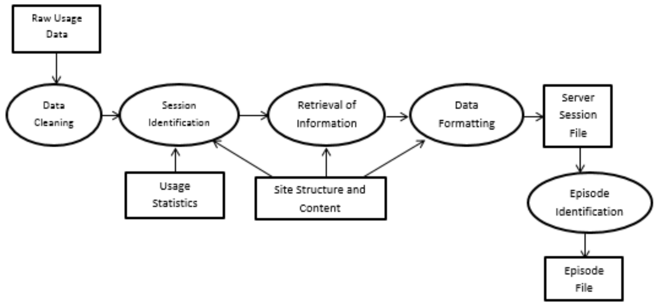

Fig. 4: Preprocessing of Web Usage Mining

Web Usage Mining techniques are being widely used in the commercial applications making it a popular choice for e-commerce. These techniques include the three main models – (i)Association Rules, (ii)Sequential Patterns and (iii)Clustering.

Association Rules is the most used technique in Web Usage Mining. It is used to find associations among web pages that frequently appear together in users' sessions.

Sequential Patterns are used to search for recurrent subsequences in the midst of large amount of sequential data. In web usage mining, sequential patterns are used to find sequential navigation patterns that appear in users' sessions frequently. Another technique of Web Usage Mining is clustering that searches for groups of similar objects among large amount of data based on a general idea of distance function which computes the similarity between groups. Web Usage Mining uses this technique extensively to group together analogous sessions [11].

4. APPLICATIONS

Web Mining has become very popular in commercial applications and is very much in demand in specific areas like e-commerce and e-business. The e-commerce and e-business also runs efficiently with the applications like text mining and data mining but web mining is considered to be best among them [11]. Few applications of web mining are given below [12]:

1. E-Commerce: Web Mining generates individual user's profile to understand the needs of users. It checks for fraud. Helps in internet advertising and also provides retrieval of similar images.
2. Information Retrieval: Search engines on the web use this application of web mining to generate topic hierarchies. Also, it is used to extract schemas for XML documents.
3. Digital Libraries: Web Mining provides us the privilege to get access to all the different books in different parts of the world at one place without being physically present there.
4. Network Management: Network Management helps to deliver the content to users reliably in a short duration of time. This is done by traffic management and fault management.

5. A FUTURE DIMENSION: CLOUD MINING

As a metaphor for the Internet, "the cloud" is a familiar term. Hence in simple words cloud computing is basically "a type of internet based computing". Cloud computing allows large groups of servers in different areas to be networked to permit the centralized storage of data and online access to computer services or resources. The use of cloud computing is becoming increasingly popular day by day due to its low cost, mobility and availability.

Cloud mining is the retrieval of resources over the cloud. In future the amounts of data will outsize the static data that cannot be stored in computer systems. To compensate the lack of storage size, data is stored over the internet. It can be used anywhere just an internet connection is required. Sharing can also be introduced with cloud data and the sharing rights are defined for a particular user.

6. CONCLUSION

It is believed that the web is growing as rapidly as numbers of sites are being added every day. With the enormous amount of data that is being added to the World Wide Web every day, it is extremely important to analyze the data patterns for the benefit of the users so that more useful web application may be developed. Web mining is an important area of research which outlines the techniques for extracting useful information from the web about user's data usage patterns. In this paper, we have discussed various techniques used for web mining. We have outlined the difference between data mining and web mining giving an overview of web content mining, web structure mining and web usage mining. Furthermore, we have also discussed its applications areas such as in digital libraries, e-commerce, information retrieval and network management.

REFERENCES

[1] Kavita Sharma, Gulshan Shrivastava and Vikas Kumar.”Web Mining: Today and Tomorrow” in Proceedings of 2011 3rd International Conference on Electronics Computer Technology (*ICECT 2011)*.

[2] Kosala, Raymond; Hendrik Blockeel "Web Mining Research: A Survey". SIGKDD Explorations (*July 2000*).

[3] Jaideep Srivastava, Web Mining: Accomplishments & Future Directions.

http://www.ieee.org.ar/downloads/Srivastava-tut-paper.pdf

[4] P.Sujath1, G.Thailambal and R.Sheela Angalin Ruby. “Study of Web Content Mining and Its Tools.” International Journal of Engineering & Computer Science *Vol.3,Issue 8 (August-2014)*

[5] Miguel Gomes da Costa Júnior and Zhingo Gong. “Web Structure Mining: An Introduction” in proceedings of the 2005 IEEE International Conference on Information Acquisition *June 27-July 3, 2005, Hong Kong and Macau, China*

[6] Zakaria Suliman Zubi “Ranking WebPages Using Web Structure Mining Concepts” Recent Advances in Telecommunications, Signals and Systems

[7] Dillip Kumar Sharma and A.K.Sharma” A Comparative Analysis of Web Page Ranking Algorithms” in (IJCSE) International Journal on Computer Science and Engineering *Vol. 02, No. 08, 2010, 2670-2676*

[8] L. Getoor. “Link Mining: A New Data Mining Challenge. SIGKDD Explorations, vol. 4, issue 2, 2003”

[9] Han, J. Kamber, M. Kamber . “Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers” *2000*

[10] Federico Michele Facca and Pier Luca Lanzi. “Recent Developments in Web Usage Mining Research” in proceedings of Data Warehousing and Knowledge Discovery *Volume 2737, 2003, Pg 140-150*

[11] S. Yadav, K.Ahmad, J.Shekar. “Analysis of Web Mining Applications and Beneficial Areas” in proceedings of the IIUM Engineering Journal, *Volume 12, no. 2, 2011*

[12] Navneet Goyal. “Web Mining” http://www.slideworld.com/slideshow.aspx/WEB-MINING-Prof-Navneet-Goyal-BITS-Pilani-ppt-681474

[13] Bamshad Mobasher. “Web Usage Mining- An introduction. http://www.powershow.com/view/9dc82-NzhmZ/Chapter_12_Web_Usage_Mining_An_introduction_powerpoint_ppt_presentation