Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 4, April 2013, pg.164 – 169

REVIEW ARTICLE

Study on Various Web Mining Functionalities using Web Log Files

Supinder Singh[1], Sukhpreet Kaur[2]

[1]Student of masters of technology Computer Science, Department of Computer Science Engineering,
Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India
[2]Assistant Professor, Department of Computer Science and Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India

[1] Supindersingh89@yahoo.com; [2] Preetsukhpreet@gmail.com

Abstract— As the size of web increases along with number of users, it is very much essential for the website owners to better understand their customers so that they can provide better service, and also enhance the quality of the website. To achieve this they depend on the web access log files. The web access log files can be mined to extract interesting pattern so that the user behavior can be understood. This paper presents an overview of web usage mining and also provides a survey of functionalities that are associated with web log files used for web usage mining.

Key Terms: - Web Usage mining; Functionalities; log files; Preprocessing

I. INTRODUCTION [1]

Web Usage Mining:

In this world of Information Technology, accessing information is the most frequent task. Every day we have to go through several kind of information that we need and what we do? Just browse the web and the desired information is with us on a single click. Today, internet is playing such a vital role in our everyday life that it is very difficult to survive without it. The World Wide Web (WWW) has influenced a lot to both users (visitors) as well as the web site owners. The web site owners are able to reach to all the targeted audience nationally and internationally. They are open to their customer 24X7. On the other side visitors are also availing those facilities.

In the last fifteen years, the growth in number of web sites and visitors to those web sites has increased exponentially. The number of users by June 30 2010 was 1,966,514,816[18] which are 28.7% of the world's population. The number of active web sites is *125,522,259* [19] as on 13-Dec-2010. Due to this growth a huge quantity of web data has been generated.

To mine the interesting data from this huge pool, data mining techniques can be applied. But the web data is unstructured or semi structured. So we cannot apply the data mining techniques directly. Rather another discipline is evolved called web mining which can be applied to web data. Web mining is used to discover interest patterns which can be applied to many real world problems like improving web sites, better understanding the visitor's behavior, product recommendation etc.

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents/services (Etzioni,1996). Web mining is categorized into 3 types.

1. Content Mining (Examines the content of web pages as well as results of web Searching)
2. Structure Mining (Exploiting Hyperlink Structure)
3. Web Usage Mining (analyzing user web navigation)

Web usage mining is a process of picking up information from user how to use web sites. Web content mining is a process of picking up information from texts, images and other contents. Web structure mining is a process of picking up information from linkages of web pages. Web Mining Content Mining Structure Mining Usage Mining

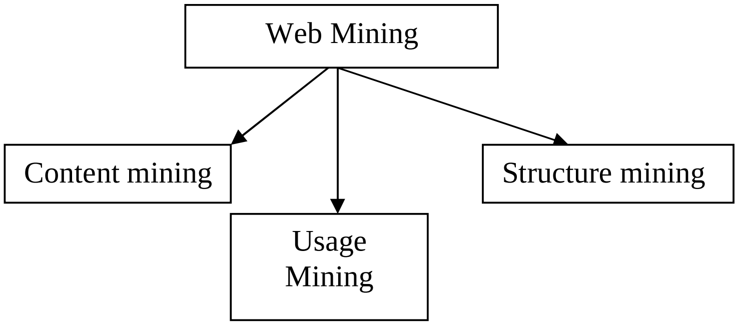

Figure 1: Web mining classification

Web Mining Functionalities:

PERSONALIZATION FUNCTIONS

A Web personalization system can order a variety of functions starting from simple user salutation o more complicated functionality such as personalized content delivery. Kobsa et al. (2001) recommends cassations of the Web personalization functions, which is extended here to a generic cassations scheme. The proposed scheme takes into account what is currently opened by commercial systems and research prototypes, as well as what is potentially feasible by such systems. We distinguish between four basic classes of personalization functions: memorization, guidance, customization and task performance support. Each of these is examined in more detail below.

Memorization:

This is the simplest form of personalization function, where the system records and stores in its ‘memory information about the user, such as name and browsing history. When the user returns to the site, this information is used as a reminder of the user’s past behavior, without further processing. Memorization, is usually not a stand-alone function, but as part of a more complete personalization solution. Examples of this class of functions are the following:

User Salutation:

The Web personalization system recognizes the returning user and displays the users name together with a welcome message. Various commercial sites employ salutation for their customers or registered users. Though this is a simple function, it is the rest step towards increased visitor loyalty, since users feel more comfortable with Web sites that recognize them as individuals, rather than regular visitors.

Bookmarking:

The system stores the Web pages that a user has visited in the past and presents them to the user by means of a personalized bookmarking schema for that site.

Personalized access rights:

A Web site can use personalized access rights, in order to separate authorized users from common users. Divergent access rights may be required for deferent types of information, such as reports or product prices, or even for the execution of particular Web applications, such as ftp, or e-mail.

Guidance:

Guidance as a personalization function refers to the endeavor of the personalization system to assist the user in getting quickly to the information that the user is seeking in a site, as well as to provide the user with alternative browsing options. This personalization function not only increases the users’ loyalty but also alleviates in a great extent the information overload problem that the users of a large Web site may face. Examples of guidance functionality are the following:

Recommendation of hyperlinks:

This function refers to the recommendation of a set of hyperlinks that are related to the interests and preferences of the user. The presentation of the recommended links is done either in a separate frame of the Web page or in a pop-up window. In (Kobsa et al.,2 001), this function is described as adaptive recommendation and can take the form of recommendation of links to certain products, topics of information, or navigation paths that a user might follow. Recommendation of hyperlinks is one of the most commonly offered Web personalization functions, and is supported by a number of systems such as the Web Personalize (Mobasher et al.,2000 b).

User tutoring:

This functionality borrows the basic notion of Adaptive Educational Systems, and applies it to Web sites. A personalized site can offer guidance to an individual at each step of the users interaction with the site, according to the users knowledge and interests. This is achieved by either recommending other Web pages, or by adding explanatory content to the Web pages. An application of this function can be found in Webinars (Web seminars), which are live or replayed multimedia presentations conducted from a Web site.

Customization

Customization as a personalization function refers to the modification of the Web page in terms of content, structure and layout, in order to take into account the user's knowledge, preferences and interests. The main goal is the management of the information load, through the facilitation of the user's interaction with the site. Examples of this class are:

Personalized layout:

This is a functionality inherited from Adaptive User Interfaces, where a particular Web page changes its layout, color, or the locale information, based on the parole of the user. This function is usually exploited by Web portals, such as Yahoo and Altavista, which are obeying customized features in order to create personalized My Portal sites.

Content Customization:

The content of the Web page presented to a user may be modified in order to adjust to the user's knowledge, interests, and preferences. For example, the same page can be presented to deterrent users, in a summarized, or an extended form depending on the type of the user. An example of such a customized Web site is the UM2001 site (Schwarzkopf,2001).

Customization of hyperlinks:

Customization can also apply to the hyperlinks within a page. In this case the site is muddied by adding or removing hyperlinks within a particular page. This can lead to the optimization of the whole Web site structure by removing links that are unusable and modifying the site's topology to make it more usable.

Personalized pricing scheme:

The Web site can provide deferent prices and payment methods to deferent users, such as discounts or installments to users that have been recognized by the site as loyal customers. An attempt of providing functionality similar to that was performed by amazon.com, which charged deferent customers with deferent prices for the same product. However, the attempt was legally challenged, due to the failure of communicating and justifying the reasons behind the price deference's. Together with hyperlink recommendation, this functionality can also be employed by e-commerce sites to attract visitors that are not currently buyers.

Personalized product differentiation:

In marketing terms, personalization can be a powerful method of transforming a standard product into a specialized solution for an individual.

Task Performance Support:

Task performance support is a functionality that involves the execution of a particular action on behalf of a user. This is the most advanced personalization function, inherited from a category of Adaptive Systems known as personal assistants (Mitchell et al., 1994), which can be considered as client-side personalization systems. The same functionality can be envisaged for the personalization system employed by a Web server.

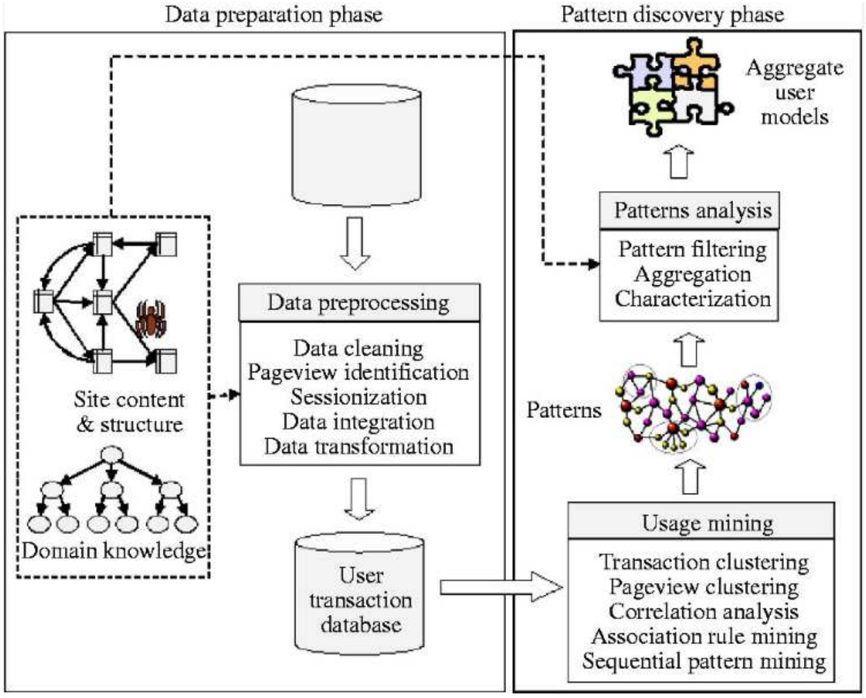

Figure 2. Represents Functionality of web log files with web Usage Mining [6]

Steps followed in web usage mining [7]

1. Data collection – Web log files, which keeps track of visits of all the visitors
2. Data Integration – Integrate multiple log files into a single file
3. Data preprocessing – Cleaning and structuring data to prepare for pattern extraction
4. Pattern extraction – Extracting interesting patterns
5. Pattern analysis and visualization – Analyze the extracted pattern
6. Pattern applications – Apply the pattern in real world problems.

II. LOG FILE [11]

.A **log file** is a recording of everything that goes in and out of a particular server. It is a concept much like the black box of an airplane that records everything going on with the plane in the event of a problem. The information is frequently recorded chronologically, and is located in the root directory, or occasionally in a secondary folder, depending on how it is set up with the server. The only person who has regular access to the log files of a server is the server administrator, and a log file is generally password protected, so that the server administrator has a record of everyone and everything that wants to look at the log files for a specific server. Servers are not the only system that use log files. Process control systems, as well as computer operating systems have logging subsystems that work exactly like a log file does. While these are more sophisticated than a simple log file, most times it is the same concept, where a log message is recorded in the file and saved until it is needed. Other forms of log filing use more sophisticated systems, some of which even analyze the logs before they are needed, but it all depends on where the log file is located. The point of a **log file** is to keep track of what is happening with the server. If something should malfunction within a complex system, there may be no other way of identifying the problem. Log files are also used to keep track of complex systems, so that when a problem does occur, it is easy to pinpoint and fix. Log files are also important to keeping track of applications that have little to no human interaction, such as server applications. There are times when log files are too difficult to read or make sense of, and it is then that log file analysis is necessary. Log file analysis is generally performed by some kind of computer program that makes the log file information more concise and readable format. Log files can also be used to correlate data between servers, and find common problems between different systems that might need one major solution to repair them all.

Log file types
Access Log, Agent Log, Error Log and Referrer Log.

Referrer log file contains the information about there for. As someone jumps from any side to www.google.com by clicking the link, referrer log file of goggle server will record a referrer entry that a user came from that particular web site.

Error log file records the errors of web site especially when user clicks on particular link and link does not locate the promised page or web site and user receives ‘‘Error 404 File Not Found’’. Error Log file is more helpful for the web page designer to optimize the web site links.

Agent log file records the information about the web site users’ browser, browser’s version and operating system. This information is again utilized by the web site designer and administrator for the analysis that users are using which specific browser to access the web site. There are number of browser available to users and each browser has its own properties and advantages to their users. Different version of same browser can different added utilities and benefits to its users, so web site can be modified accordingly. Information about the users’ operating system is also help for designer and web site changes are made accordingly.

Access Log File is major log of web server which records all the clicks, hits and accesses made by any web site user. There are number of attributes in which information is captured about users. Information about the user is then processed for WUM and user behavior and interest can be mined. Table 1 elaborates the different attributes of access log file along with their description. There are three main types of web server log file formats available to capture the activities of user on web site. All the three log files are in ASCII text format. Log files act as health monitor for the web sites and are main source of user access data and user feedbacks. These are Common Log File Format (NCSA); Extended Log Format (W3C); and IIS Log Format (Microsoft).

NCSA Common log file format is most widely used to capture user data. It is standardized format but not customizable. Only fixed numbers of attributes are available for raw data of users. Figure 1 elaborates the example of common log file with basic necessary information of log entries.

III. Conclusion

We have surveyed different functionalities of web Usage Mining which are related with web files also. These web log files records information of each user request. Advantage of log files - data is easily available to be analyzed. Log files usually contain noisy and ambiguous data. Preprocessing involves removal of unnecessary data from log files. Log file used for debugging purpose. Data preprocessing is an important steps to filter and organize appropriate information before using to web mining algorithm. This paper present review for web usage mining with field extraction and data cleaning. Preprocessing web log file is used in data mining techniques, also can be used in intrusion detection system as input to detect intrusion. Hence our approach will focus on the efficient retrieval of information from web log file than before in Future Work.

References

[1] Sachin Pardeshi and Ujwala Patil “Central web mining services – public and free access log files” Proceedings of National Conference on Emerging Trends in Computer Technology (NCETCT-2012) Held at R.C.Patel Institute of Technology, Shirpur, Dist. Dhule, Maharashtra,India. April 21, 2012

[2] Priyanka Patil and Ujwala Patil “Preprocessing of web server log file for web mining” Proceedings of National Conference on Emerging Trends in Computer Technology (NCETCT-2012) Held at R.C.Patel Institute of Technology, Shirpur, Dist. Dhule, Maharashtra,India. April 21, 2012

[3] Shiv Pratap Singh Kushwah, Keshav Rawat, Pradeep Gupta “Analysis and Comparison of Efficient Techniques of Clustering Algorithms in Data Mining” International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-1, Issue-3, August 2012

[4] Liu Kewen, “Analysis of Preprocessing Methods for Web Usage Data”,1ntemational Conference on Measurement, Information and Control (MIC),2012

[5] Neelamadhab Padhy, Dr. Pragnyaban Mishra , and Rasmita Panigrahi. “The Survey of Data Mining Applications and Feature Scope”, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012.

[6] Amit Sharma, S.N. Panda and Ashu Gupta, “Data Mining Techniques and their role in Intrusion Detection

Systems", J. Acad. Indus. Res. Vol. 1(4) September 2012

[7] S.K.Pani, L.Panigrahy, V.H.Sankar, Bikram Keshari Ratha, A.K.Mandal, S.K.Pandhi "Web Usage Mining: A Survey on Pattern Extraction from Web Logs" International Journal of Instrumentation, Control & Automation (IJICA), Volume 1, Issue 1, 2011

[8] Mrs. G. Sudhamathy, Dr. C. Jothi Venkateswaran, "Web Log Clustering Approaches – A Survey", G. Sudhamathy et al. / International Journal on Computer Science and Engineering (IJCSE) 7 July 2011

[9] Theint Theint Aye, "Web Log Cleaning for Mining of Web Usage Patterns" ,IEEE 2011

[10] Gary M. Weiss, Brian D. Davison, "Data Mining" To appear in the Handbook of Technology Management, H. Bidgoli (Ed.), John Wiley and Sons, 2010

[11] Tasawar Hussain, Dr. Sohail Asghar, Dr. Nayyer Masood tasawar, "Web Usage Mining: A Survey on Preprocessing of Web Log File",2010.

[12] Juan Vel´asquez, Hiroshi Yasuda and Terumasa Aoki, "Combining the web content and usage mining to understand the visitor behavior in a web site",Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03),IEEE 2003.