International Journal of Allied Practice, Research and Review
Website: www.ijaprr.com (ISSN 2350-1294)

A Study of Processes Involved in Web Usage Mining

Neeraj Kandpal[1], Prof. Ripu Ranjan Sinha[2], M. S. Shekhawat[3]
[1]Research Scholar, Suresh gyan vihar University, Jaipur, Rajasthan.
[2]Prof.(Research), Suresh gyan vihar University, Jaipur, Rajasthan.
[3]Department of Physics, Govt. Engineering College, Bikaner, Rajasthan.

Abstract: Web mining is mining the log data for various modern applications. Web mining is divided into three types web content mining, web structure mining and web usage mining. Web usage mining is the data mining technique to mine the web log data from World Wide Web to extract useful information. Web usage mining useful for the applications like personalized marketing, fraud detection, to identify criminal activities, web design etc. This paper is going to explain in detail about the process involved in Web Usage Mining.

Keywords: Web Usage Mining, Web log files, Web usage mining process, Web usage mining.

I. Introduction

Web mining is the application of data mining techniques to extract knowledge from Web data including Web documents, text on web, images on web, usage logs of web sites etc. Web Mining can be categorized into three broad areas of mining[1]. Web Content Mining (WCM) is mining of information from the contents of web like text, images etc. Web Structure Mining (WSM) deals with the structure of the web pages and effect of this structure on traversal through web pages. Web usage mining is a research field that focuses on the development of techniques and tools to study users web navigation behavior [1]. Web Usage Mining is also called web log mining. Web Usage Mining is the application of data mining techniques to discover useful usage patterns from Web data, in order to understand and better serve the needs of Web based applications. The main applications of Web Usage Mining are personalization of web content, computational Intelligent combinations, web design, ecommerce, Web advertising/marketing, pre-fetching and caching, transaction Analysis, modification of web site, fraud detection, customer relationship management, product/Site recommendation, identify web robots, to improve web server program's performance etc.

When web user makes a request to web server, all information about users visit will be recorded in web log server files. A log file resides at Web Servers, Web proxy Servers and Client browsers [2].Web Server Log files resides in web server and collect all activities of users browsing website. There are four types of web server logs i.e., access logs, agent logs, error logs and referrer logs. Information of proxy server resides in web proxy server log files. Client browser Log Files resides in client's browser and to store them special software are used.

Some tools used to explore Web Usage Mining are Web Utilization Miner(WUM), KOINOTITES, Web miner, Web Site Information Filter System(Web SIFT), WebViz, WET (Web-Event Logging Technique), WebQuilt, Web log miner, Web mate, Web usage miner, i-miner.

II. Research Process

Process involved in Web Usage Mining

The steps involved in Web Usage Mining are as follows:

(A) Web usage data Collection and Integration: This step involves extraction of log data from server log files. Data can be basically collected from three sources [1,12,13]:

1. *Server side data collection:* These logs files usually contain basic information e.g.: name and IP of the remote host, date and time of the request, the request line exactly as it came from the client, etc. This information is usually represented in standard format.

2. *Proxy Side data collection:* The main difference with the server side is that proxy servers collect data of groups of users accessing groups of web servers.

3. *Client Side data collection:* Accessed data can be achieved by using various modern techniques.

Integrate multiple log files into a single file is defined as data integration.

(B) Preprocessing of data: Preprocessing phase is very important step of web usage mining [4]. The main steps of preprocessing are:

1. *Data Cleaning/ Data Reduction:* In this step voluminous data that is obtained from servers cleaned by using some criteria, by applying an effective algorithm. The purpose of data reduction process is to remove unwanted data that may affect the overall mining process[3]. Helmy et al. [5] removed any extensions like gif, jpg, css in target URL. Use of this algorithm, these types of useless data is removed and the mining process gets be evaluated results comparatively fast. The HTTP status code is also a concern for data reduction. In the web personalization area a researcher take the entire data log that contains success code 200 series. As in web intrusion detection, all the status code of server errors is most important because in successful status code an almost no margin to find a suspect. Suneetha et al. [6] give details of HTTP status code. In anomaly user behavior investigation the failure error i.e. 400 series code and server error 500 series code is important. So in the web log entries that contains 400 and 500 status code is not eliminated in data reduction phase.

2. *User Identification:* According to Chaofeng [7], each IP address represents one user. An IP address represents a different user, if a page is requested by a referrer link; there is another user with the same IP address. Cooley et al. [8] proposed a heuristic that if a web page is directly accessed without any hyperlink by same IP, assumed as a different user. User

identification refers to identify unique users. Users with different ip address are considered unique users.

3. *Session Identification:* Some researchers [3] have coined that there is a new session if time limit is exceed more that 30 minutes. Session identification defined by differentiates the web log entries into different user sessions by a session timeout. Once a user was identified then click stream is divided into clusters. This method of division is called Session Reconstruction or Sessionization.

4. *Path Completion:* Path completion used to check the missing pages after constructing transactions. The missing page problem is due to proxy servers and caching problem of clients.

(C) Pattern discovery of web usage data: In pattern discovery phase three main operations Association, Clustering and Sequential Analysis, are performed on data for pattern discovery [11].

a. *Association analysis:* It works on generating frequent pattern and rules. In web log file number of URL visit by number of users so we can identify frequently accessed web pages by users which can help to understand user needs. The association rule is mainly focuses on discovery of relations between pages visited by users on web site. Association rule can be used to relate the web page is most often used by the single server session. Several algorithms like Apriori, Eclat, Frequent Pattern tree etc. to perform association rule mining.

b. *Clustering Analysis:* Clustering is unsupervised learning technique. Clustering analysis defined as similar characteristics users are group together without knowledge of group definition. Clustering will help us to find group of common behavior users. Clustering of web pages is very important for internet service provider to analyze the behavior of users. Clustering can also be used for anomaly detection. Data's which are not fit to any type are not related to any cluster. These cases are anomalies.

c. *Sequential pattern analysis:* There are several algorithms like AprioriAll, GSP, SPADE, PrefixSpan and Spam are used for sequential pattern analysis. This analysis is used to find that a suspected user visit a particular link X followed by link Y in a time ordered set of sessions [10]. By using this approach we can predict the suspected user psychology which is useful in crime detection. There are several algorithms like AprioriAll, GSP, SPADE, PrefixSpan and Spam are used for sequential pattern analysis.

d. Classification: Server data is classified according to one or more common attributes. It makes the bunches of similar type of records.

(D) Pattern Analysis of web usage data: In pattern analysis step we find out a valuable model or standard pattern for specific web usage mining application. Techniques used for pattern analysis are visualization technique, OLAP techniques, data and knowledge querying and usability analysis [3, 9].

a. *OLAP (Online Analytical Processing Technique):* Sales, marketing, management reporting, business process management, financial reporting etc are some of the applications of *OLAP (Online Analytical Processing Technique)*. *It* is a powerful paradigm for strategic analysis of relational database which is very useful in business systems [3]. Business intelligence, relational reporting, data mining and many more are related to OLAP.

b. *Data and Knowledge Querying:* SQL is the most common method of pattern analysis. This is an important part of web usage mining in which we analyze the different reasons of anomaly behaviors of users. By the use of SQL we find some specific results from database like suspected session in database created by the users like failure status code of http protocol in very short interval of time.

c. *Usability analysis* is a modeling technique to accessing the behavior of user on the web site.

d. *Visualization Technique:* The behavior of web users can be effectively represented by graphical method. Graphical method is representation of visualization technique.

III. Conclusion

Detailed study has been conducted for processes involved in web usage mining. Preprocessing is very important step in web usage mining for reducing large data set. Pattern discovery has been improved in recent days due to the advancement in search engine algorithms. Similarly pattern analysis has been improved by sophisticated and high speed computers. Web usage mining is power tool for standardization of web practices. It is useful for the digital forensics investigations, transaction identification, crime investigation, automated website design modification and many more applications related to web server mining.

IV. References

[1] Bharti Joshi, Ph.D., Suhasini Parvatikar, "Analysis of User Behavior through Web Usage Mining", International Journal of Computer Applications, International Conference on Advances in Science and Technology (ICAST-2014).

[2] K. R. Suneetha, Dr. R. Krishnamoorthi, " Identifying User Behavior by Analyzing Web Server Access Log File" IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009.

[3] Amit Pratap Singh , Dr. R. C. Jain, "A Survey on Different Phases of Web Usage Mining for Anomaly User Behavior Investigation", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 3, Issue 3, May – June 2014.

[4] C. Sakthipriya , G. Srinaganya , Dr. J. G. R. Sathiaseelan, "An Analysis of Recent Trends and Challenges in Web Usage Mining Applications", International Journal of Computer Science and Mobile Computing, Vol. 4, Issue. 4, April 2015, pg.41 – 48.

[5]Mohd Helmy, Abd Wahab and Nik Shahidah, "Development of Web usage Mining Tools to Analyze the Web Server Logs using Artificial Intelligence Techniques", The 2nd National Intelligence Systems and Information Technology Symposium (ISITS 207), October 30-31 2007, ITMA-UPM, Malaysia.

[6] K. R. Sunnetha and Dr. R. Krishnamoorthi, "Identifying User by Analyzing Web Server Access Log File", International Journal of Computer Science and Network Security (IJCSNS), Vol. 9, No. 4, 2009, pp. 327-332.

[7]Li. Chaofeng, " Research and Development of Data Preprocessing in Web Usage Mining", International Conference on Management and Engineering, 2006, pp. 1311-1315.

[8]R. Cooley, B. Mobasher and J. Srivastava, "Data Preparation for Mining World Wide Web Browsing Patterns", Journal of Knowledge and Information Systems, Springer, 1999, Vol. 1, No. 1, pp. 1-27.

[9] Aarti M. Parekh, Anjali S. Patel, Sonal J. Parmar, Prof.Vaishali R. Patel, "Web usage Mining:Frequent Pattern Generation using Association Rule Mining and Clustering", International Journal of Engineering Research & Technology (IJERT), Vol. 4 Issue 04, April-2015.

[10] D. Antoniou, M. Paschou, E. Sourla, and A. Tsakalidis, "A Semantic Web Personalizing Technique The case of bursts in web visits," presented at IEEE Fourth International Conference on Semantic Computing, 2010.

[11] Khushbu Patel, Anurag Punde, Kavita Namdev, Rudra Gupta, Mohit Vyas, DETAILED STUDY OF WEB MINING APPROACHES-A SURVEY, INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY, Patel, 4(2): February, 2015.

[12] V.Chitraa,Dr. Antony Selvdoss Davamani, "A Survey on Preprocessing Methods for Web Usage Data", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 3, 2010.

[13] Vijayashri Losarwar, Dr. Madhuri Joshi , "Data Preprocessing in Web Usage Mining", International Conference on Artificial Intelligence and Embedded Systems , 2012 .