A Survey of Issues and Techniques of Web Usage Mining

Preeti Rathi[1], Dr (Mrs) Nipur Singh[2]

Research Scholar, Dept. of Computer Science, Kanya Gurukul Campus, Dehradun, India,
Professor, Dept. of Computer Science, Kanya Gurukul Campus, Dehradun, India,

Abstract-*In era of technology mining play an important role in field of computer science. The World Wide Web is an interactive and popular platform to transfer information. Web Usage Mining is the type of web mining and it is application of data mining techniques. Web Usage Mining has become helpful for website management, personalisation etc. Usage data internment the origin of web users along with their browsing behaviour at a website. It means weblog records to discover user access pattern from web pages. Weblog contains all information regarding to users which is useful to access pattern. Web mining helps to gather the information from customer who's visiting the site. Now a days various issues related to log files i.e. data cleaning, session identification, user identification etc. In this survey paper we discuss the phases of WUM, architecture of WUM, issues related to WUM and also discuss the future direction.*

Key Words: Data Pre-Processing, Intrusion data, Log files, Web Usage Mining, data cleaning.

1. INTRODUCTION

With the brisk growth of the World Wide Web, the web has become an imperative medium of information dissemination. Therefore, the information available on the Web has become a vital source of information for the users of the internet. When data mining techniques are applied to Web data, it is referred to as Web mining. In 1996 its Etzioni **[4]** was first to coin the term web mining. For example in various websites contains various webpages and webpages having various pattern, through web mining we extract useful pattern from websites. According to analysis targets, web mining can be divided into three different types [**9]**

- ❖ Web usage mining
- ❖ Web content mining
- ❖ Web structure mining.

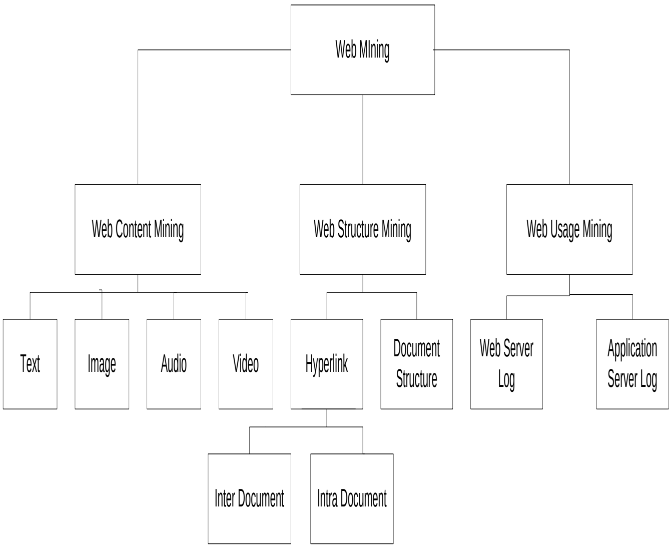

Fig-1: Taxonomy of Web Mining

1.1 Web Content Mining

In this mining extraction and integration of data, information from the content of web page. Web content mining also known as text mining. Web content mining provides the results lists to search engine in order of highest relevance to the keywords in query. Content data corresponds to the collection of facts a Web page was designed to convey to the users. Web content may be unstructured (plaintext), semi- structured (HTML documents), or structured (extracted from databases into dynamic Web pages).

1.2 Web Structure Mining

Web structure mining provide relationships between linked. The main purpose for structure mining is to extract previously unknown relationship between web pages. According to the type of web structural data, web structure mining can be divided into two kinds:

1. Extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects the web page to a different location.

2. Mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage.

1.3 Web Usage Mining

WUM is the main area of my research. Basically we work on log files. Web usage mining is the process of extracting useful information from server logs. Some users might be looking at only textual data, whereas some others might be interested in multimedia data. WUM help to find the behaviour of user and according to behaviour personalize the websites. Log files contain “what happened when by whom” i.e. log files are primary source of data. There are two types of log files- Common log file, extended log file.

2. PHASES OF WEB USAGE MINING

Web mining process can be regarded as a three-phase process consisting:

2.1 Pre-processing / Data preparation

Weblog data are pre-processed in order to clean the data moves log entries that are not needed for the mining process, data integration, identify users, sessions, and so on.

2.2 Pattern discovery

Statistical methods as well as data mining methods (path analysis, Association rule, Sequential patterns, and cluster and classification rules) are applied in order to detect interesting patterns.

2.3 Pattern analysis

Discovered patterns are analysed here using OLAP tools, knowledge query management mechanism and intelligent agent to filter out the uninteresting rules/patterns.

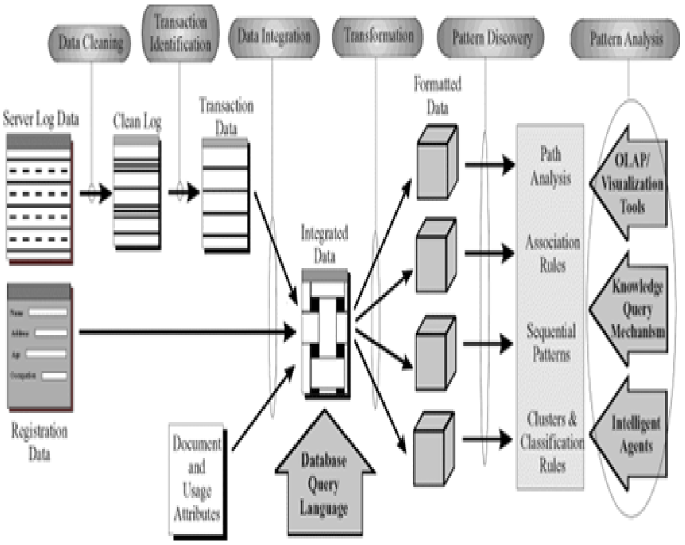

Fig-2: Architecture of Web Usage Mining

3. DATA SOURCES OF WEB USAGE MINING

Mainly there are four types of data sources present in which usage data is recorded at different levels they are: client level collection, browser level collection, server level collection and proxy level collection. **[2]**

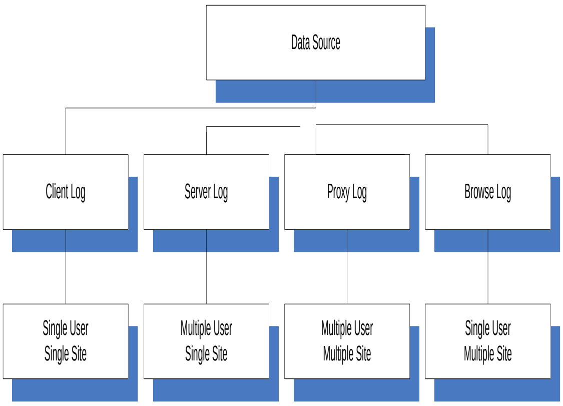
Data Source
Client Log
Server Log
Proxy Log
Browse Log
Single User Single Site
Multiple User Single Site
Multiple User Multiple Site
Single User Multiple Site

Fig - 3: Classification of Log Files

4. RELATED WORK

For getting till here there were several literatures which were helping hand. Several techniques of web usage mining like data cleaning, user identification, path completion, session identification etc. There are various related work done in log files or pre-processing of data but there are various problems are occur. Gajendra Singh, Priyanka Dixit **[14]** describe the how web log mining facing large amount of log data it is produce through web usage pattern and behaviour of user using web mining techniques. Mitali Srivastava, Rakhi Garg, P K Mishra **[15]** classify data pre-processing is considered as an important phase of Web usage mining due to unstructured, heterogeneous and noisy nature of log data. Complete and effective data pre-processing insures the efficiency and scalability of algorithms used in pattern discovery phase of Web usage mining. Kumar **[7]** propose an effective and enhanced data pre-processing methodology which produces an efficient usage patterns and reduces the size of weblog and enhance the performance of webpages. Chhavi Rana et al **[12]** presented the requirement of present scenario and some challenges and issues of future direction to enhance the quality of website. Dr. S. Vijiyarani **[16]** studied the basic concepts of web mining, classification, processes and issues. Vaishali A.Zilpe, Dr. Mohammad Atique **[17]** describe Web service providers want to find the way to predict the users behaviours and personalize information to reduce the traffic load and design the Web-site suited for the different group of users. **[6]** Describes the format and types of log files which is maintained by the web servers and also discuss the techniques of WUM. By analysing these log files gives a neat idea about the user. M. Gomati **[5]** Web mining plays an important role in discovering such knowledge, it is roughly divided into three categories: Web Content Mining, Web Usage Mining and Web Structure Mining. FST (Fuzzy Set Theory) is used to handle such data. S. K. Pani, L. Panigrahy, V.H.Sankar, Bikram Keshari Ratha, A.K.Mandal, and S.K.Padhi **[10]** presents an overview of web usage mining and also provides a survey of the pattern extraction algorithms used for web usage mining. Bhupendra Kumar Malviya **[8]** describe the web usage mining research tools and problem related to this areas. Brijendra Singh, Hemant Kumar Singh **[13]** describe past, current evaluation and update in each of the three different types of web mining i.e. web content mining, web structure mining and web usages mining. These following table summarize the latest work on various algorithm-

Table-1: Previous Related Works on Various Algorithm

Author's Name	**Algorithm Used**	**Year**	**Techniques**	**Description**	**Parameter Used**
Gajendra Singh Chandel , Kailash Patidar, Man Singh Mali **[18]**	FCM , K-Mean	2016	Clustering	In this paper author gives the FCM algorithm and comparison experimental result to previous algorithm K-Mean and show the FCM result is accurate in previous one and using Visual basic 6.0 used as a backend and Ms-access as frontend for implementation.	In this paper author used following parameters in FCM algorithm i.e. cluster centre, accuracy.
Ruchika R. Patil, Amreen Khan **[19]**	Bisecting K-Mean (combination of K-Mean, Hierarchical clustering)	2015	Clustering	In this paper author gives the brief idea of previous k-Mean algorithm and derived a new algorithm BKM it is combination of K-Mean & Hierarchical clustering and improve accuracy and reduce the time and show experimental result comparison to K-Mean algorithm.	In this paper following parameters are used in BKM algorithm i.e. time, accuracy.
M. Santhana Kumar, C. Christopher Columbus **[20]**	FCM , K-Median, DBSCAN , K-Mean Clustering	2015	Clustering	In this paper author used Rapid Miner tool for experimental comparison of algorithm FCM and K-Mean & calculation based on Euclidean Distance.	In this paper used parameter such as Euclidean Distance, Patterns.
Gajendra Singh, Priyanka Dixit**[14]**	Apriori Algorithm	2014	Web Log Miner (Log Files)	In this paper author used Web Log Miner for mine the log files and reduced the irrelevant data and compare the result with previous result and reduced execution time and increase CPU utilization.	In this paper parameters are Execution time, Throughput, CPU Utilization
R.Shanthi, Dr.S.P.Rajagopalan **[21]**	Apriori-All Algorithm, Web log mining Algorithm	2013	Log Files, Navigation Pattern	In this paper author comparison Apriori algorithm to new algorithm Apriori All algorithm is modification of previous algorithm based on time and join.	Users and User ID used as a parameter in Apriori All algorithm
A. K. Santra, S. Jayasudha **[22]**	Naïve Bayesian	2012	Classification	In this paper Naïve Bayesian algorithm for applied on weblog and reduced the time to find the result and compare result with previous result	In this paper for classification Weblog parameter is used and calculate the result.

				C4.5 decision tree algorithm. This algorithm is used for personalization.	

5. ISSUES IN WEB USAGE MINING

There are various issues related to WUM. These issues are identifying users(User Identification) **[11]** it is major issue because many user have multiple address and user or we can say client have one address then identify the user problem occur because logical entity used to identify the user. Missing Data **[3]** it is another problem arise in log files because some time user used backward and forward button so information should be lost. Noisy Data **[3]** Noisy data is corrupt data or we can say that un-useful data, and this type of data is not relevant to the client. Intrusion Data **[1]** Intrusion means any set of action threatens the integrity & availability & confidentiality of network resources. So we can say that intrusive data is another issue in web usage mining. Session Identification **[11]** it is also called access sequence. In session identification identify or finding the significant accessing sequence.

6. CONCLUSION & FUTURE DIRECTION

This paper has discussed about the Web Mining and its types, and also discuss the issues related to log files. In this paper we conclude that WUM is mining process to extract useful pattern from log files and enhance the performance of web pages using personalization. In future direction we check user behaviour to determine whether data is error free or not.

REFERENCES

[1] Masoud Najjar Barghi, "An Effective WebMining-based Approach to Improve the Detection of Alerts in Intrusion Detection Systems", International Journal of Advanced Computer Science and Information Technology (IJACSIT), Vol. 4, No. 1, 2015, Page: 38-45, ISSN: 2296-1739.

[2] Sanjeev Dhawan, Swati Goel," Web Usage Mining: Finding Usage Patterns from Web Logs" American International Journal of Research in Science, Technology, Engineering & Mathematics, 2013, Page: 203-207, ISSN (Print): 2328-3491, ISSN (Online): 2328-3580, ISSN (CD-ROM): 2328-3629.

[3] Hassan F. Eldirdiery, A. H. Ahmed,"Detecting and Removing Noisy Data on Web Document using Text Density Approach", International Journal of Computer Applications (0975 – 8887), Vol. 112, No. 5, February 2015, Page: 32-36.

[4] Oren Etzioni," The World-Wide Web: Quagmire or Gold Mine?" ACM, Vol. 39, No. 11, November 1996, Page: 66-68.

[5] M. Gomati ," A Survey on Web Mining Using Fuzzy Logic", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 5, Issue 3, March 2015, ISSN: 2277 128X.

[6] L.K. Joshila Grace, V.Maheswari, Dhinaharan Nagamalai, "Analysis of Web Logs and Web User in Web Mining", International Journal of Network Security & Its Applications (IJNSA), Vol. 3, No.1, January 2011, Page: 99-110.

[7] M.Praveen Kumar," An Effective Analysis of Weblog Files to improve Website Performance", International Journal of Computer Science & Communication Networks, Vol. 2(1), Page: 55-60, 2011, ISSN: 2249-5789.

[8] Bhupendra Kumar Malviya, Jitendra Agrawal, "A Study on Web Usage Mining: Theory and Applications", Fifth International Conference on Communication Systems and Network Technologies, IEEE, Page: 935-939, April 2015, ISBN (Print) 978-1-4799-1797-6/15

[9] R.Natarajan, Dr.R.Sugumar, "A Survey on Attacks in Web Usage Mining" International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 5, Page: 4470-4475, May 2014, ISSN (Online): 2320-9801, ISSN (Print): 2320-9798.

[10] S. K. Pani, L. Panigrahy, V.H.Sankar, Bikram Keshari Ratha, A.K.Mandal, S.K.Padhi, "Web Usage Mining: A Survey on Pattern Extraction from Web Logs", International Journal of Instrumentation, Control & Automation (IJICA), Vol. 1, Issue 1 , 2011, Page:15-23.

[11] Sheetal A. Raiyani, Rakesh Pandey, Shivkumar Singh Tomar, "Performance Enhancement of Web Server log for Distinct User Identification through different Factors", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 6, June 2014, Page: 7262-7267, ISSN (Online) : 2278-1021, ISSN (Print) : 2319-5940.

[12] Chhavi Rana, "A Study of Web Usage Mining Research Tools", Int. J. Advanced Networking and Applications, Vol. 3, Issue: 06, Pages: 422-1429, 2012, ISSN: 0975-0290.

[13] Brijendra Singh, Hemant Kumar Singh, "Web data Mining Research: A Survey", Computational Intelligence and computing Research, IEEE, December 2010, Page: 1-10, Print ISBN: 978-1-4244-5965-0.

[14] Gajendra Singh, Priyanka Dixit, "A New Algorithm for Web Log Mining" International Journal of Computer Applications (0975 – 8887) Vol. 90, No. 17, March 2014, Page: 20-24.

[15] Mitali Srivastava, Rakhi Garg, P K Mishra, "Analysis of Data Extraction and Data Cleaning in Web Usage Mining", ICARCSET, Unnao, India, March 2015, ISBN: 978-1-4503-3441-9.

[16] Dr. S.Vijiyarani, Ms E. Suganya, "Research Issues in Web Mining", International Journal of Computer-Aided Technologies (IJCAx), Vol. 2, No.3, July 2015, Page: 55-64.

[17] Vaishali A.Zilpe, Dr. Mohammad Atique, "Neural Network Approach for Web Usage Mining", National Conference on Emerging Trends in Computer Science and Information Technology (ETCSIT), Page: 31-33, 2011.

[18] Gajendra Singh Chandel, Kailash Patidar, Man Singh Mali ," A Result Evolution Approach for Web usage mining using Fuzzy C-Mean Clustering Algorithm", IJCSNS International Journal of Computer Science and Network Security, Vol.16, No.1, January 2016, Page: 135-140.

[19] Ruchika R. Patil, Amreen Khan," Bisecting K-Means for Clustering Web Log data", International Journal of Computer Applications (0975 – 8887), Vol. 116, No. 19, April 2015, Page: 36-41.

[20] M. Santhana Kumar, C. Christopher Columbus, "Web Usage Based Analysis of Web Pages Using Rapid Miner", Wseas Transactions on Computers, Vol. 14, 2015, Page- 455-464, E-ISSN: 2224-2872.

[21] R.Shanthi, Dr.S.P.Rajagopalan ,"An Efficient Web Mining Algorithm To Mine Web Log Information " , International Journal of Advanced Research in Computer Science and Software Engineering , Vol. 1, Issue 7, September 2013, ISSN(Online): 2320-9801.

[22] A. K. Santra, S. Jayasudha, "Classification of Web Log Data to Identify Interested Users Using Naive Bayesian Classification", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No. 2, January 2012, Page: 381-387, ISSN (Online): 1694-0814.