Journal Of Harmonized Research (JOHR)

Journal Of Harmonized Research in Engineering
2(4), 2014, 424-432

ISSN 2347 – 7393

Review Article

RESEARCH TOOLS & TECHNIQUES OF WEB USAGE MINING

J. Haweliya

Institute of Engineering & Technology,
DAVV, Indore (M.P.), India

Abstract*:* - World Wide Web is a huge repository of web pages. Understanding the web user behavior is the key of success for any web based business or application. Analysis of web user behavior helps webmaster to provide good web navigation experience of the website. So, web is the abundant area of Data Mining research. Web mining is the extension of Data Mining. Web Usage Mining is the area of Web mining which emphasis on taking out of interesting knowledge from Web logs produced by Web servers. Cloud mining can be seen as a future of Web mining. This paper gives the categorization of web mining (i.e. Web content mining, Web structure mining, and web usage mining). This paper focuses on the various research tools and techniques of Web Usage Mining.
***Keywords*—** Web Usage Mining tool; Web Content Mining; Web Structure Mining; Web Usage Mining;

I. Introduction:

With the growth of technology at a faster rate, World Wide Web has also grown-up exponentially. Now a day it extends to whole world and being used by all fields. World Wide Web become universal and acts as a very powerful tool to collect share and broadcast information along the world. As the popularity of web increased, its complexity also increases due to the existence of bulk amount of data. In day-to-day life web turns into a huge repository of information and it become very complex for the user to retrieve/extract the useful information from such a huge nugget. Web mining is an interesting research area which combines the both Data Mining and World Wide Web. Oren Etzioni was the first person who coin the term Web mining in his paper. Etzioni starts by making hypothesize that the information on the Web is adequately structured and also outlines the subtasks of Web mining [1]. In this paper he describes the Web mining processes. We can state that the web mining is as the discovering and analyzing process toward retrieve the useful/meaningful information from the www. Although there is so much research have been done on data mining on the web but

For Correspondence:
jyoti.samad@rediffmail.com
Received on: November 2014
Accepted after revision: December 2014

still researcher have the scope in this area. Mining of data is basically varied from structured to unstructured. The primary focus of Data mining is on structured data that is organized in a database while the text mining focuses on unstructured data. Web mining deals with the semi-structured and/or unstructured data. In the process of Web mining web data triggers more complexity because it is available in unstructured form. Getting the information from web has become a very challenging task. The techniques of Web mining behaves like a device to carry out this challenge. These techniques are very helpful in automatic discovery and retrieval of information from the internet [2].

The Web mining is categorized into three: Web Content mining, Web Structural mining and Web Usage mining [3]. Pulling out of useful content from the structured and/or unstructured web document is explained in Web Content Mining [4]. The addressing troubles of Web search and automatic community detection has been solved by two algorithms that work on the Web graph [5]. A complete framework and observations to retrieve the useful patterns from log files of a real Web site has been described in [6].

The structure of this paper is as follows: Section 2 describes the taxonomy of Web mining in which it encompasses a brief description of different types of Web mining. Section 3 describes various techniques of Web usage mining. Section 4 deals with the literature survey which produce a brief description of the current researches done in the area of web usage mining. Section 5 briefly describes the various tools used in web usage mining and also their features, section 6 describes the conclusion of the article and section 7 is about the references used.

II. Taxonomy of Web Mining

An application of data mining is web mining to pull out knowledge from web data which includes web documents, hyperlinks between documents, usage logs of web sites, etc. The concentration paid to web mining is on research, software industry, and web based organization, has led to the accumulation of significant experience. Some researchers suggest that the task of web mining is disintegrating into following sub tasks [7-8]:

- *Resource Discovery:* - It locates he unfamiliar documents and services on the Web.
- *Information Selection and Pre-Processing:-* In this process automatic extraction and pre-processing of specific information is done from newly discovered Web resources.
- *Generalization:* - It uncovers the general patterns at individual Web sites and across multiple Sites.
- *Analysis:* - In the analysis the Validation and interpretation of mined patterns is done.
- *Visualization:* - It represents the results of an interactive analysis in a visual as well as easy to understand fashion.

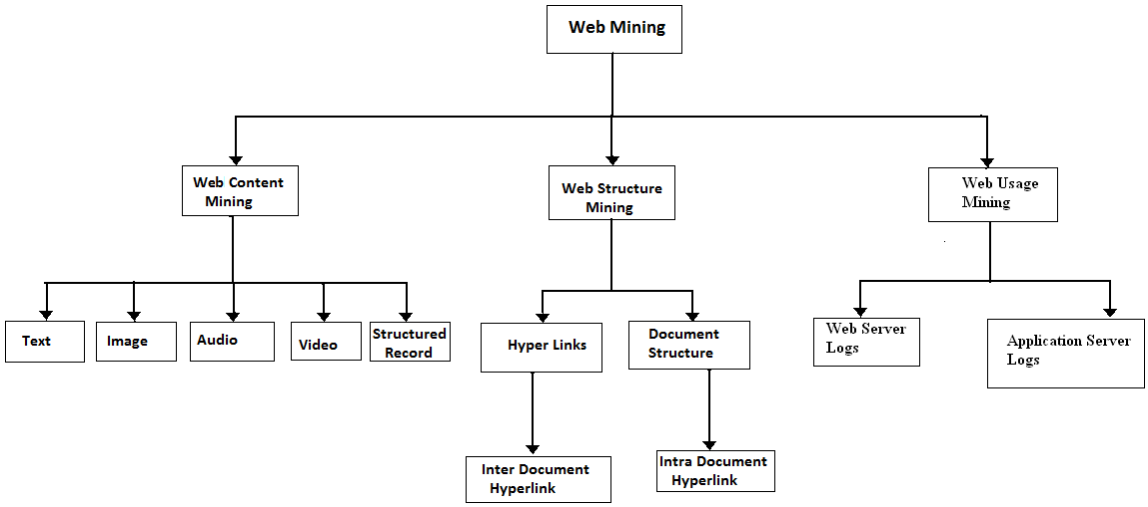

Fig. 1: Taxonomy of Web Mining

Broadly Web mining is divided into three categories. They are:-

A. Web Content Mining: - It is the process to retrieve useful information from the contents of web documents. Content data have the collection of facts a Web page was designed to convey to the users. It consists of text, audio, video, images or structured records such as lists and tables [9].

B. Web Structure Mining: - Web Structure mining can be defined as a process of extracting the structure information from the web. The structure mining is based on Web graph. As a normal graph Web graph also contains nodes and edges. Here Web pages works as nodes and hyperlinks works as edges which connects one page to other page if they are related. This mining can be performed either at the document level (i.e. intra page) or at the hyperlink level (i.e. inter page). Web Structure mining is used to improve structural design of websites.

C. Web Usage Mining: - It is a process to discover the meaningful patterns from data generated by client-server transactions on one or more Web localities. Here the mining is performing on the usage pattern of the data that is stored in web logs. It mainly deals with getting information about the navigational pattern of the user. Some examples of web logs which generated automatically are accesslogs, referrerlogs and client-side cookies.

III. **Techniques of Web Usage Mining**

Web Usage mining is a center of attention due to the techniques that could predict user behavior when the user interacts with the Web. As the exponential growth of the internet Web is converted into a huge amount of data repository. These data could range very widely but mainly we classify them into the usage data that reside in the Web server, Application server and proxy servers [10]. The process of Web Usage mining can be decomposed into three-phase. It consists of the data preprocessing tasks, pattern discovery and pattern analysis phases [11] (See figure 2). Figure 2 also give a complete idea about the research areas of Web usage mining. In the first phase, Web log data are preprocessed in order to identify users, sessions, page views, and so on. In the second phase, statistical methods, path analysis as well as data mining methods (such as association rules, sequential pattern discovery, clustering, and classification) are applied in order to detect interesting patterns. In the third/last phase stored patterns could be further analyzed in the process of Web usage mining [12]. Web Mining uses several techniques of data mining to retrieve the useful facts from internet. Except the data mining techniques some other techniques are also available such as artificial intelligence, natural language processing, information retrieval, machine learning, information extraction which can also be applied.

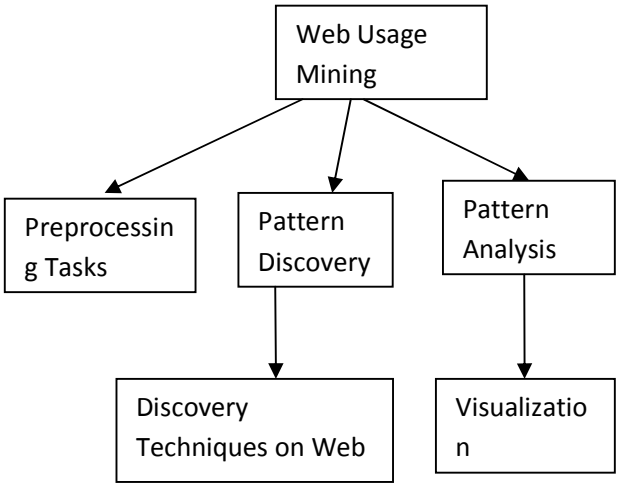

Fig. 2: Area of Web Usage Mining

There are so many techniques used for various applications. The major application areas of web usage mining are: personalization, System improvements, Site modification, Business intelligence and Usage characterization. The table1 briefly describes the various techniques used in different phases of Web Usage mining.

Table1: Various Mechanisms used in Web Usage mining

Phase of Web Usage Mining	Used Mechanism	Description
Preprocessing Tasks	Usage Preprocessing	Data cleaning, session reconstruction, content and structural information retrieval and data abstraction.[13]
	Content Preprocessing	
	Structure Preprocessing	
Pattern Discovery	Statistical Analysis	Frequency, mean, and median are determined on sessions and results of this analysis represent the most frequently accessed web pages, mean view time and/or length of the each page [14].
	Association Rules	Find out the correlations among all web pages that frequently arise concurrently in a user browsing session (Apriori Algo. is the most common) [15].
	Clustering	The main objective of clustering mechanism in Web Usage mining is to assemble the similar sessions together. (k-means with genetic algo., EM-CFuzzy means algo etc. are used) [16].
	Classification	It is supervised way of learning to associate the data items with one of many predefined classes. (Naïve Bayes classifier, SVM etc. are used)[17].
	Sequential Pattern	It is used to discover the sessions those are found in a chronological order. They include the series of items which frequently occur in a particular order (MIDAS algo. most commonly used) [18].
Pattern Analysis	Knowledge Query Mechanism	It is used to extract the useful patterns from discovered Pattern (SQL is mainly used).
	OLAP/Visualization Technique	OLAP provides a framework for analysis that allows changes in aggregate levels. The output of OLAP queries will be an input for data mining or data visualization tools.
	Intelligent Agents	Various intelligent agents help to examine the patterns in web usage mining. These intelligent agents accomplish the work of analyzing the discovered patterns.

IV. Literature Survey:

There have been some works about content mining, and structure mining, Data mining (research based), Information Extraction, and Artificial Intelligence. The web usage mining research area, several groups did significant work. Exact Web usage information could help to engage new customers, retain current customers, exceed cross marketing/sales, track leaving customers and discover the most effective logical structure for their Web space [19]. The comprehensive surveys on web Usage mining have been done by Koutri Avouris,, and Daskalaki [20]. Pierrakos et al. [21] Kosala and Blockeel [22] research on the terms of web mining and the related area earlier in their work. An intelligent multi-agent based environment well known as intelligent Java Development Environment (iJADE) to serves an integrated and Intelligent agent based platform in the e-commerce environment on Internet shopping proposed by Lee and Liu [23]. The purpose of this application is intelligent agent for helping users is applied in

various applications and not only in e-commerce environment. Mobasher et al. [24] states that effective and scalable techniques for Web personalization rooted on association rule discovery by Usage data. Toolan and Kushmerick [25] proposed various kind of methods based on web usage mining to deliver personalized Site Maps that are specialized to the interest of each individual visitor. Improved the performance of web mining compared to traditional approach such as database approach by applying the agent technology. Web usage mining is being used in numerous regions. In [26] web usage mining is used for Enhance the scalability and answer time of search engines. A simple storage model signifies that main memory is assigned with the function of doing dynamic caching of the answers and lists that are present in secondary memory. Lu, Dunham, and Meng [27] later proposed a technique to initiate remarkable Usage Patterns (SUP) and used them to obtain significant "user preferred navigational trails". Falkowski et al. [28] proposed two methods to inspect the evolution of two different types of online communities. A new tool for web usage mining that depend on the bio-mimetic relational clustering algorithm based on computation to produce an efficient visualization of the activity of users on a website is introduced by Labroche, Lesot, and Yaffi [29]. Khalil, Li and Wang [30] published an improved Web page prediction accuracy by applying a novel approach that consists integrating clustering, association rules, and Markov models. This integration offers superior prediction accuracy than using each technique separately. [31] An approach for tracking evolving user profiles and enriches it with straightforward information gathered from web log data. With this profiles are also provides other domain specific information and a validation procedure is implied to estimate the caliber of the mined profiles. Data sources called intentional browsing data (IBD) for potentially upgrade the effectiveness of WUM applications explore by Tao, Hong and Su [32]. David et al. [33] proposed a probabilistic model for a web site that uses the entropy of a Markov chain in sequence to compute the user navigation patterns from the log data. Most of the research in the area of web usage mining focus on the algorithm while disgarding the type of data on which the algorithm will be concern. Masseglia et al. [34] proposed to perform a separate data mining process to express frequent behaviors by discovering the den sest periods. This period are the one having at least one frequent sequential pattern? Basically this set for the users connected to the Web site in that particular period. A review has been done on developments in web usage mining research [35]. The process, techniques and applications are discussed of web usage mining. Dai and Mobasher [36] explained the need to associate Web usage and content knowledge, with the help of enhancing the information in the Web usage logs with semantics acquire of the Web site's pages. Thakare and Gawali [37] emphasized on the importance of the effective and preprocessing of access stream ahead actual mining process can be done, that could significantly upgrade the automatic discovery of meaningful pattern and interaction from access stream of user. An algorithm is based on association rule mining with the help of incremental techniques proposed by Rao, Kumari, and Raju, [38].They explain rule mining with incremental technique to suit the dynamically changing log scenario which is more systematic that running a number of scans of database. A comparison done by Kumar and Rukmani [39] how A priori algorithm and Frequent Pattern Growth algorithm differs. Comparison in terms of memory and time usage while discovering the web usage patterns of Websites. In the form of frequent sequences outcome of semantic knowledge by the patterns produce for Web Usage mining explained by Senkul and Salin [40] .These continual navigational patterns are composed by ontology instances instead of Web page addresses.

V. Tools Used in Web Usage Mining

There are various tools used in Web Usage mining. Each tool has different functionality. It is selected by the user according to their application. The list of these tools with their description and URL is given in the table 2.

Table 2: Various tools used in Web Usage mining

S. N.	Tools	Description	URL
1.	Data Preparator	It is a free software tool designed to assist with common tasks of data preparation (such as cleaning, extraction and transformation).	http://www.datapreparator.com
2.	Lisp Miner	It performs data preprocessing by analyzing the click stream and data collected.	http://lispminer.vse.cz
3.	SpeedTracer	This tool helps to identify and fix performance problems in your web applications. It Mines web server logs and reconstruct the user navigational path for session identification.	https://developers.google.com/web-toolkit/speedtracer
4.	i-Miner	It is used to discover data cluster by using fuzzy clustering algorithm and fuzzy inference system for pattern discovery and analysis.	http://www.solutionmetrics.com.au/products/iminer
5.	Rapid Miner	It is an advanced analytics, including predictive analytics, data mining, and text mining. It automatically and intelligently analyzes data – including databases and text – on a large scale.	http://rapidminer.com
6.	MiDas(Mining Internet Data for Associative Sequence	It discovers marketing based navigational pattern from log files. It applies more features to traditional sequential method.	http://www.midasplatform.org/MIDAS/resources/toolbox.html
7.	Angoss Knowledge WebMiner	It Combines ANGOSS Knowledge STUDIO and clickstream analysis	http://www.angoss.com
8.	ClickTracks	By this tool Visitor patterns can be shown on Web site.	http://www.clicktracks.com/products/optimizer/index.php
9.	Megaputer WebAnalyst	The tool has data and text mining capabilities.	http://www.megaputer.com/site/index.php
10.	MicroStrategy Web Traffic Analysis Module	Traffic highlights, content analysis, and Web visitor analysis reports can be generated.	http://www2.microstrategy.com/bi-applications/solutions/website-analytics.asp
11.	SAS Web Analytics	It is used to analyze Web site traffic.	sas.com/solutions/webanalytics
12.	SPSS Web Mining for Clementine	Mainly used for extraction of Web events.	http://www-01.ibm.com/software/analytics/spss
13.	WebTrends	It is used to give Web traffic information.	webtrends.com
14.	XML Miner	System and class library for mining data and text expressed in XML, using fuzzy logic expert system rules	http://www.xml.com
15.	STATISTICA Data Miner	This software provided by statSoft which can process, red and write data from all file format.	https://www.statsoft.com/Products/STATISTICA/Data-Miner
16.	Web log Expert	It is a fast and powerful access log analyzer which can generate reports in HTML, PDF and CSV formats.	http://www.weblogexpert.com
17.	KXEN Modeler	It provides all the functions of data mining such as regression, classification, clustering, attribute importance, segmentation forecasting and association rule data.	http://www.sap.com/pc/analytics/predictive-analytics/software/infiniteinsight/index.html
18.	WEKA Tool	This software is provided by university of Waikato. It provides functions like data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.	http://www.cs.waikato.ac.nz/ml/weka

IV. Conclusion

This paper expresses the Web as a communication and information medium by a majority of the population. Web data is growing at a significant rate. So many new Computer Science concepts and techniques have been developed. This paper gives the fertile area of research in Web Usage mining. Now a day Web Usage mining is used in so many areas like e-marketing, e-education, e-commerce, bioinformatics and digital libraries etc. Statistical Analysis, clustering, association rules etc. are the various technique by which Usage patterns are discovered. Efficiency of pattern analysis is basically improved by involving the intelligent agents and knowledge query mechanisms. This paper also describes various research tools used in the area of Web usage mining.

REFERENCES

[1] O. Etzioni, "The World-Wide Web: quagmire or gold mine", Communications of the ACM, 39(11), pp 65-68, 1996.

[2] M. Eirinaki and M. Vazirgiannis.,"Web Mining for web Personalization.", ACM Trans. Inter. Tech. Vol. 3, No.1, pp 1-27, 2003.

[3] B. Singh, H.K. Singh, "WEB DATA MINING RESEARCH: A SURVEY", Computational Intelligence and Computing Research (ICCIC), IEEE International Conference, 2010.

[4] N. Anwat, V. Patil, "Survey Paper on Web Usage Mining for Web Personalization", International Journal Of Innovative Research & Development, vol. 3, Issue 7, pp 127-132, 2014.

[5] M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, Andrew S. Tomkins, "The Web as a Graph: Measurements, Models, and Methods", SpringerLink, Vol. 1627, pp 1-17, 1999.

[6] O. Nasraoui, M. Soliman, E. Saka, A. Badia, R. Jermain, " A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites ", IEEE, 2008.

[7] Q. 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 pp. 683–720, 2008.

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

[9] J. Shrivastava, "Web Mining :Accomplishments & Future Directions", University of Minnesota, USA, 2002.

[10] J. Srivastava, R. Cooley, M. Deshpande, P. Tan, "Web Usage mining: Discovery and applications of usage patterns from web data" SIGKDD Explorations newsletter, 1(2), pp 12-23, 2000.

[11] B. Mobasher, R. Cooley, and J. Srivastava, J., "Automatic personalization based on Web usage mining", Communications of the ACM, 43(8), 142-151, 2000.

[12] M. Eirinaki and M. Vazirgiannis, "Web mining for web Personalization", ACM Transactions on Internet Technology, 3(1), pp 1-27, 2000.

[13] T.T. Aye, Univ. of Comput. Studies, Mandalay, Mandalay, Myanmar, Computer Research and Development (ICCRD), 3rd International Conference on (Vol.: 2), 2011.

[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] Yanxi Liu ; Sch. of Sci., Changchun Univ., Changchun, China, "Study on Application of Apriori Algorithm in Data Mining", Computer Modeling and Simulation, ICCMS '10. Second International Conference on (Vol.:3), 2010.

[16] Supreet Kaur1, Usvir Kaur2, "A Survey on Various Clustering Techniques with K-means Clustering Algorithm in Detail ",International Journal of Computer Science and Mobile Computing, Vol. 2, Issue. 4, pp155 – 159, 2013.

[17] A.G.Buchner, M. Baumgarten, S.S.Anand MD.Mulvenna and J.G.

Hughes."Navigation Pattern Discovery from Internet Data", In WEBKDD, San Diego, CA 1999.

[18] R. Cooley,"Web Usage Mining: discovery and Applications of Interesting patterns from web data", Ph.D. Thesis, University of Minnesota, 2000.

[19] J. I. Hong, , J. Heer, S. Waterson, and J. A. Landay , "Web Quilt: A proxy-based approach to remote web usability testing", ACM Transactions on Information Systems, 19(3), pp 263-285, 2001.

[20] M. Koutri, N. Avouris, and S. Daskalaki," A survey on web usage mining techniques for web-based adaptive hypermedia systems", Adaptable andAdaptive Hypermedia Systems Idea, pp 1-23, 2004,.

[21] D. Pierrakos, G. Paliouras, C. Papatheodorou, and C.D. Spyropoulos, "Web usage mining as a tool for personalization: A survey", User Modeling and User Adapted Interaction, 13(4), pp 311-372, 2003.

[22] R. Kosala, and H. Blockeel, Web Mining Research: A Survey', Machine Learning, 2(1), pp 1-15, 2000.

[23] R. S. T. Lee, and J. N. K. Liu, iJADE, "Web-Miner: An Intelligent Agent Framework for Internet Shopping", IEEE Transactions on Knowledge and Data Engineering, 16(4), pp 461- 473, 2004.

[24] B. Mobasher, H. Dai, T., Luo, and M. Nakagawa, "Effective personalization based on association rule discovery from web usage data", Proceeding of the third international workshop on Web information and data management WIDM 01, 9, USA, pp 9-15, 2001.

[25] F. Toolan, and N. Kusmerick, "Mining Web Logs for Personalized Site Maps", Proceedings of the Third International Conference on Web Information Systems Engineering (Workshops) - (WISEw'02 (WISEW '02). IEEE Computer Society, Washington, DC, USA, pp 232-237, 2002.

[26] J. Kerkhofs, K. Vanhoof and D. Pannemans,"Web Usage Mining on Proxy servers: A Case Study", Limburg University Center, July 30, 2001.

[27] L. Lu, M. H. Dunham, and Y. Meng, "Discovery of Significant Usage Patterns from Clusters of Clickstream Data", Proceedings of the ACM SIGKDD workshop on Knowledge Discovery in Web WebKDD05, Chicago, IL, USA, 2005.

[28] T. Falkowski, J. Bartelheimer, and M. Spiliopoulou," Mining and Visualizing the Evolution of Subgroups in Social Networks", Proceedings of the 2006 IEEEWICACM International Conference on Web Intelligence, pp 52-58, 2006.

[29] N. Labroche, M. J. Lesot, and L. Yaffi, "A New Web Usage Mining and Visualization Tool", 19th IEEE International Conference on Tools with Artificial Intelligence ICTAI 2007, 1, pp 321-328, 2007.

[30] F. Khalil, J. Li, and H. Wang, "Integrating Recommendation Models for Improved Web Page Prediction Accuracy", Reproduction, 74(Acsc), Australian Computer Society, Inc. ACM International Conference Proceeding Series, Wollongong, Australia, Vol. 312, pp 91-100, 2008.

[31] O. Nasraoui, M. Soliman, E. Saka, A. Badia, and R. Germain, "A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites", IEEE Transactions on Knowledge and Data Engineering, 20(2), pp 202-215, 2008.

[32] Y. Tao, T. Hong, and Y. Su, "Web usage mining with intentional browsing data, Expert Systems with Applications", 34(3), pp 1893-1904, 2008.

[33] N. David, L. Patrascu, A. Sasu, & D. Damian, "A probabilistic model for web usage mining" , Proceedings of the 8th Wseas international conference on Telecommunications and informatics, World Scientific and Engineering Academy and Society (WSEAS), pp 129–133, 2009.

[34] F. Masseglia, P. Poncelet, M. Teisseire, and A. Marascu, "Web usage mining: extracting unexpected periods from web logs", Data Mining and Knowledge Discovery, 16(1), pp 39-65, 2008.

[35] G.R.B., S.Totad, P.PVGD.,"Amalgamation of Web Usage Mining and Web Structure Mining.", International Journal of Recent trends in Engineering Vol.1, Issue 2, 2009.

[36] H. Dai, and B. Mobasher," Integrating Semantic Knowledge with Web Usage Mining for Personalization", Information Systems Journal, pp 1-28, 2009.

[37] S. B. Thakare, and S. Z. Gawali, "A Effective And Complete Preprocessing For Web Usage Mining", International Journal On Computer Science And Engineering, 2(3), pp 848-851, 2010,.

[38] M. Rao, M. Kumari, and K. Raju, Understanding User Behavior using Web Usage Mining', International Journal of Computer Applications, 1(7), pp 55–61, 2010.

[39] B. S. Kumar, and K. V. Rukmani, "Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms", International Journal of Advanced Networking and applicatons, 1(06), pp 400-404, 2010.

[40] P. Senkul, and S. Salin, "Improving Pattern Quality in Web Usage Mining by Using Semantic Information", Knowledge and Information Systems, 1(6), pp 400- 404, 2011.

[41] K. Chaudhary, S. K. Gupta, "Web Usage Mining Tools & Techniques: A Survey", International Journal of Scientific & Engineering Research, Vol. 4, Issue 6, pp 1762-1768, 2013.