Volume 7, Issue 5, May 2017 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Qualitative Evolution of Performance and Validity Indices for Web Usage Mining

[1]Narendra Kumar Kachhwaha, [2]Bhupendra K Malviya
[1]M.Tech (CSE), Research Scholar
[1, 2] Department of Computer Science & Engineering, School of Research and Technology, People's University, Bhopal, Madhya Pradesh, India

Abstract*— World Wide Web is an enormous store of links and web pages. It provides huge amount of information for the Internet clients. The development of web is great as about one million pages are added every day. Users' accesses are traced in web logs. Because of the great usage of web, the web log files are increasing at a more rapidly rate and the range is becoming enormous. Web Usage Mining relates mining techniques in log data to extract the performance of users which is used in different applications like Support to the Design, E-commerce, Modified services, pre-fetching etc. Web usage mining has three phases as pre-processing, pattern detection and pattern learning. Web log data is generally noisy and confusing, thus pre-processing and pattern analysis is an essential method before mining. For learning patterns gathering are to be constructed professionally. This paper is presents work done in the web usage mining. Finally a glance of various applications of web usage mining is presented. Web Usage Mining has develop into a dynamic region of study in field of data mining because of its crucial values. This paper affords a widespread conversation of the all the stages in Web Usage Mining and Problems with related works in this research areas.

Keywords*— Web Usage Mining, Data mining, Web Mining, Data Preprocessing, Pattern detection, and Learning Pattern.

I. INTRODUCTION

In this planet of Information expertise, accessing information is the mainly recurrent task. Day after day we include going through various types of information that we require and what we achieve? Presently surf the web and the preferred information is through us on just a single click. Nowadays, World Wide Web or internet is playing such an essential responsibility in our daily living life that it is extremely complicated to live without it. The internet has partial loads of to both users (clients) and the web site vendors. The website vendors are capable to achieve to all the intentioned users all over the country and globally. They are serving to their clients 24 hours. On the other part clients are also gaining those services [1].Information in Web Usage Mining can be achieved in server logs, proxy logs, browser logs, and simultaneously from a group's database. This information set fluctuate in requisites of the site of the data resource, the class of data existing, the area of people on or after which data was acquired, or method of execution [1].Web Usage Mining is a branch of Web Mining, which is a division of Data Mining. The method of mining considerable and precious data or information from enormous database is named Data Mining. Web Usage Mining extracts (mines) the usage attributes of the clients of Web services. This achieved records can then be useful in a different approaches for example, inspection of false essentials etc [2].Web Usage Mining is well thought-out as a constituents of the big business Intelligence in a business. It is apply for choosing business advances through the proficient use of Web services. It is incredibly critical for the Customer Relationship Management (CRM) because it assurance clients performance till the interface among the organization and the client is concerned [3].

II. WEB USAGE MINING

Web log mining is also known as web usage mining is the application of data mining methods on huge web log storage areas to determine valuable knowledge concerning customer's behavioral patterns and website usage information to facilitate for a variety of website design fields. The key resource of statistics for web usage mining consists of contents logs together with some web servers incomparable in the world. There are three stages in web usage mining. Users log data is collectively beginning different resources as alternative servers, client side and server side so on. Data Preprocessing: makes a succession of processing of web log file covering records cleaning, session recognition, customer recognition, path finishing position and transaction detection. Pattern detection: principle of different data mining methods to processed data related to statistical study, association, clustering, and pattern subsequent and so on. Pattern learning: formerly patterns are discovered as of web logs, tedious convention are filtered away. Learning is completed with information investigation system for occurrence data cubes or SQL to presents OLAP (On-Line Analytical Processing) actions. The entire three stages are representing through the following diagram.

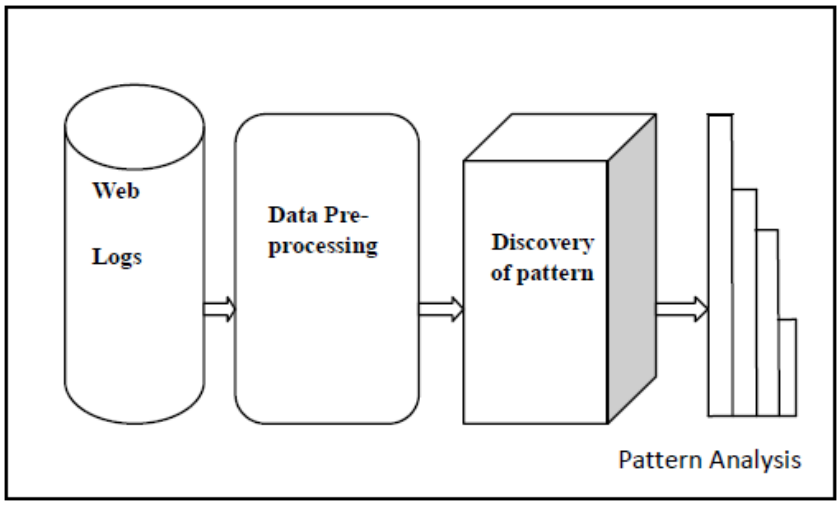

Fig. 1: Stages of Web Usage Mining

The objective of this paper is to present a study of web usage mining along with an analysis of three stages. Data gathering sector records out different data resources and the different areas of functions or applications where web usage mining is appropriate [4].

A. Noteworthy Contribution in the Field Of Web Usage Mining:

A performance of data preprocessing system for web usage mining and the details of algorithm for path achievement are presented in Yan LI's [5]. After user session recognition, the missing pages in user access paths are added by using the referrer-based technique which is an efficient solution to the problems initiated by using proxy servers and local caching. The reference length of pages in absolute path is modified by taking into consideration the average reference length of auxiliary pages which is expected in advance during the maximal forward references and the reference length algorithms. As established by practical web access log, the algorithm path achievement, proposed by Yan LI, efficiently joined the lost information and recovers the reliability of access data for auxiliary web usage mining calculations. Huiping Peng [6] used FP-growth algorithm for processing the web log records and obtained a set of frequent access patterns. Then using the combination of browse interestingness and site topology interestingness of association rules for web mining they discovered a new pattern to provide valuable data for the site construction.In order to solve some existing problems in traditional data preprocessing technology for web log mining, an improved data preprocessing technology is used by the author ling Zheng [7]. The identification strategy based on the referred web page is adopted at the stage of user identification, which is more effective than the traditional. One based on web site topology. At stage of Session Identification, the strategy based on fixed priori threshold combined with session reconstruction is initiated. First, the original session set is developed by the technique of fixed priori threshold, and then the original session set is optimized by using session rebuilding. Experiments have established that advanced data preprocessing knowledge can improve the quality of data preprocessing outcomes.In Web Usage Mining, web session clustering plays an important key role to categorize web visitors on the basis of user click history and similarity evaluate. Swarm based web session clustering assists in various ways to manage the web resources successfully such as web personalization, schema modification, website alteration and web server presentation. Doru Tanasa[8], in his study brought two large contributions for a Web Usage Mining process. First they proposed a comprehensive methodology for preprocessing the Web logs and second are a divisive universal methodology with three approaches as well as associated concrete processes for the discovery of sequential patterns through a low support. Jiang Chang-bin and Chen Li [9] brought regarding a Web log data preprocessing algorithm supported on mutual filtering. It can achieve user session identification rapidly and flexibly even though statistic records are not enough and user history visiting records are not presence. Dr. Sohail Asghar, Tasawar Hussain, [10] proposed a structure for web session clustering at preprocessing stage of web usage mining. The structure covers the data preprocessing steps to organize the web log records and exchange the categorical web log information into numerical data. A session vector was achieved, so that suitable similarity and swarm optimization might be applied to cluster the web log information. Author says that the hierarchical cluster based advance enhances the existing web session methods for more structured information regarding the user session.

III. DIFFICULTIES WITH WEB USAGE MINING

The following problems are point out in Web related study and applications:

A. Discovery of Large Information

To get particular information on the web, users mostly either surf Web documents directly or related a search engine as a search supporter. While a client operates a look for engine to locate information, he or she commonly enters one or some keywords as an uncertainty, and then the search engine lead a directory of ranked pages derived from the significance to the uncertainty. But, there are usually two most important concerns related with the query-based Web search [13].

Authors	***Source Of log File***	***Preprocessing Technique***	***Algorithm Applied***
Qinjiao MAO and Yan LI's, Boqin FENG [5]	English Study Web site Log File	Session Identification, Client Identification, Data Cleaning, Path Completion, Transaction Identification.	Maximal Forward References(MFR), Reference Span
Gui and Feng Li Ling Zheng, Hui [7]	IIS (Internet Information Server) Server Log File	Log File Cleaning, Data Cleaning, Session Identification, User Identification.	Based on referred web page and Set priori threshold
Brigitte Trousse and Doru Tanasa [8]	Log Files from INRIA web sites	Data Structuration, Data Cleaning, Data Union, Data Summarization.	NA
Chen Li JING and Chang-bin [9]	Web server Log file	Data Preprocessing.	Based on Mutual Filtering
Sohail Asghar, Tasawar Hussain, Nayyer Masood[10]	Server Log File	User Identification, Log File Filtering Data Clean-up, Session Identification.	NA
S. Chenthur Pandian and J. Vellingiri [11]	Institution Web Site	Path Completion, Data Cleaning, Transaction Identification, User Identification, Session Identification.	MFR RL & Instance Window
Huang Zhiqiu and Fang Yuankang [12]	Chizhou College Website	Session Identification, Information Filtering .	Frame page and Page Threshold

The preliminary problem is low accurateness, which is origin by many of unrelated pages returned by the search engine. The subsequently problem is low remember, which is caused by the lack of capability of indexing all Web pages presented on the Web sites. These reasons the complexity in finding the un- indexed records that is essentially significant. How to get more appropriate pages to the uncertainty, thus, is attractive a accepted issues in Web data organization in last decade [14].

B. Finding Essential Information

Many search engines performs in a query-triggered techniques that is mostly on a basis of one keyword or some keywords entered. From time to time the results returned by the search engine do not accurately match what a client actually essential due to the information of the reality of the homology [15]. For instance, when one client with an information understanding background requests to search information by respect to "Python" programming words, he or she might be reachable with information on the creatural python, a form of snake rather than the programming words, known entering only a "python" declaration as query. In other terms, the semantics of Web data is occasionally taken into explanation in the perspective of Web search [16].

C. Learning helpful knowledge

*A*long with traditional Web search verify, query results suitable to query input are arrival to Web clients in a prearranged list of pages. In different cases, we are concerned not only browsing the returned anthology of Web pages, although furthermore extracting potentially helpful knowledge beyond them. Further interestingly, learning has been performed on how to extend the Web as an information base for evaluation creation or knowledge innovation in current times [17, 18].

IV. APPLICATIONS OF WEB USAGE MINING

Web usage mining has various applications areas for example site reorganization, web link prediction, web personalization and pre-fetching. The primary two areas are concert topic. The next two uses areas can be evaluate in e-commerce context. Various other applications of web usage mining are.

A. Upgrading in System Performance:

Performance of web services is a significant problem for user satisfaction. Web usage mining is an important research area for discovering web traffic procedures, which can be used to amplify new policies for raising the web server concert [19, 20]. Load balancing, Web caching method transmission or account distributions are the general application areas of web mining for performance upgrading. Expecting the next web page request for an exacting user collection is an essential research problem in this environment. This process is valuable mainly in web services using static relaxed since dynamic relaxed decrease the usability of web caching at both user and server-level. With proxy in classify for pre-fetching a page is well another method for performance upgrading [21, 22].

B. Site restructuring

The link association and content composition of any website are two major factors any web site. The current trend in web mining tools go towards shorter navigation sequences, in support of that motive the accessibility of objective page in several web domain needs to be improved [23]. Restructuring site topology of any web field can achieved this. The restructuring task can be performing with respect to the frequent patterns unconcerned at the end of web usage mining. Web usage data gives information along with the design of any web site with interests to user's activities [24]. To show page-stay moment gives the pages, which is not attractive. Vendor of Web site can re-establish these pages and observe the behavior of users on these pages. These two organizations techniques content and structure leads to adaptive web sites. The model given in changes web site organization with respect to usage patterns uncovered [25, 26].

C. Site Personalization

Personalization of web site is one of the key issues in many web-based applications such as individual marketing like electronic import. Performing dynamic recommendation is a very important feature in many applications like cross-sale and up sales in e-commerce. Web usage mining is the basic research area, which can be an tremendous approach for this type of problems [27, 28]. Existing recommendation Systems do not use data mining for recommendations. Web site personalization is based on usage information. Web server logs can be used to cluster web users having related interests. This system contains two modules called offline and online [29]. Offline module creates clusters using log information and online module is used for dynamic link generation. All site users is assigned to single cluster based on his/her current traversal pattern. The links that are showed on browser of any user is the same with other users in the similar cluster [30, 31].

D. Supporting to the Design

The Usability is most important issues in the design and accomplishment of web sites. The consequences shaped by Web Usage Mining methods can present strategies for improving the design of web functions. The utilize stereogram estimate the association and the efficiency of web sites from the clients view-point [32]. Web Usage mining methods to propose appropriate variation for web sites. Adaptive Web sites characterize a further step. Within this case, the substance and the arrangement of the web site can be enthusiastically reorganized alongside with the data mined from the client's behavior [33].

E. E-commerce through web usage mining

Mining E-commerce intelligence from web usage data is significantly significant for web-based businesses. An effective advantage from the use of Web Usage Mining methods can have Client Relationship Management (CRM). Within this case, the aim is on business explicit problems such as: customer removal, customer attraction, cross sales, and customer retention [34] [35] [36].

V. CONCLUSION

In this paper, we discuss different research open concerns on web mining and web usage mining. We conclude here Web usage mining model is a type of mining to server logs. Web Usage Mining act an important part in realizing the usability of the website design, the improvement of customer's relations and improving the necessity of system presentation and so on. Web Usage mining presents the support for the providing personalization server, web site design and other business making decision etc. In this paper we conclude some problems with web usage mining such as finding desirable information, finding related information, learning valuable knowledge, recommendations or personalization of data, and recent work in web usage mining research field. We also compare the web usage mining with data mining and warehousing on a few factors. We observe the web usage mining is very essential now a day for web world and more improvement required in future for user privacy.

REFERENCES

[1] Han, Jiawei, Micheline Kamber, and JianPei., "Data mining: concepts and techniques", publisher: Morgan Kaufmann, 2006.

[2] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", SIGKDD Explorations, Vol. 1, No. 2, Pp. 12-23, 2000.

[3] Shailey Minocha, Nicola Millard, Lisa Dawson, "Integrating Customer Relationship Management Strategies in (B2C) E-Commerce Environments", IFIP Conference on Human-Computer Interaction- INTERACT, 2003.

[4] F. Johnson and S. K. Gupta, Web Content Mining Techniques: A Survey, International journal of computer applications (0975-888), Volume 47– No.11, June 2012.

[5] Yan LI, Boqin FENG and Qinjiao MAO, "Research on Path Completion Technique In Web Usage Mining", IEEE International Symposium On Computer Science and Computational Technology, pp. 554-559, 2008.

[6] Huiping Peng, "Discovery of Interesting Association Rules Based on Web Usage Mining", IEEE Coference, pp.272-275, 2010.

[7] Ling Zheng, Hui Gui and Feng Li, "Optimized Data Preprocessing Technology For Web Log Mining", IEEE International Conference On Computer Design and Applications (ICCDA), pp. VI-19-VI-21, 2010.

[8] Doru Tanasa and Brigitte Trousse,"Advanced Data Preprocessing for Intersites Web Usage Mining ", Published by the IEEE Computer Society, pp. 59-65, 2004.

[9] JING Chang-bin and Chen Li, " Web Log Data Preprocessing Based On Collaborative Filtering ", IEEE 2nd International Workshop On Education Technology and Computer Science, pp.118-121, 2010.

[10] Tasawar Hussain, Dr. Sohail Asghar and Nayyer Masood, "Hierarchical Sessionization at Preprocessing Level of WUM Based on Swarm Intelligence", 6th International Conference on Emerging Technologies (ICET) IEEE, pp. 21-26, 2010.

[11] J. Vellingiri and S. Chenthur Pandian, "A Novel Technique for Web Log Mining with Better Data Cleaning and Transaction Identification",Journal of Computer Science, pp. 683-689, 2011.

[12] Fang Yuankang and Huang Zhiqiu, "A Session Identification Algorithm Based on Frame Page and Pagethreshould", IEEE Conference, pp.645-647, 2010.

[13] Eirinaki,Magdalini, and Michalis Vazirgiannis, "Web mining for web personalization", ACM, no. 1, pp 1-27, 2003.

[14] F. Toolan, and N. Kusmerick, "Mining Web Logs for Personalized Site Maps", International Conference on Web Information Systems Engineering, IEEE, pp 232-237, 2002.

[15] Nasraoui, Olfa, Maha Soliman, Esin Saka, Antonio Badia, and Richard Germain, "A web usage mining framework for mining evolving user profiles in dynamic web sites", IEEE ,no. 2 ,pp 202-215, 2008.

[16] Hasan, Tahira, Sudhir P. Mudur, and Nematollaah Shiri, "A session generalization technique for improved web usage mining", ACM, pp. 23-30,2009.

[17] Zhang, Y., J.X. Yu, and J. Hou, "Web Communities: Analysis and Construction". Berlin Heidelberg: Springer, 2006.

[18] Chang, G., et al, "eds. mining the World Wide Web: An Information Search Approach". The Information Retrieval, Vol. 10, 2001.

[19] Etminani, K.; Delui, A.R.; Yanehsari, N.R.; Rouhani, M.; "Web usage mining: Discovery of the users' navigational patterns using SOM", First International Conference on Networked Digital Technologies (NDT '09), Pp. 224 – 249, 2009.

[20] Pierrakos, D., et al, "Web Community Directories: A New Approach to Web Personalization", 1st European Web Mining Forum, pp. 113-129, 2003.

[21] Mobasher, B., "Web Usage Mining and Personalization", Handbook of Internet Computing, M.P. Singh, Editor, CRC Press. p. 15.1-37, 2004.

[22] De Lucia, A.; Francese, R.; Scanniello, G.; Tortora, G.; "Reengineering Web applications based on cloned pattern analysis", Proceedings 12th IEEE International Workshop on Program Comprehension, Pp. 132 – 141, 2004.

[23] Jing Wang, Ying Liu, Yong Shi, and Xingquan Zhu, Pushing Frequency Constraint to Utility Mining Model, ICCS 2007 Springer-Verlag Berlin Heidelberg, Part III, LNCS 4489, 2007, 685-692.

[24] Kudelka, M.; Snasel, V.; Lehecka, O.; El-Qawasmeh, E.; "Semantic Analysis of Web Pages Using Web Patterns", IEEE/WIC/ACM International Conference on Web Intelligence, Pp. 329 – 333, 2006. [25] Eirinaki,Magdalini,and Michalis Vazirgiannis, "Web mining for web personalization", ACM, vol. no. 1,pp 1-27,2003.ms and Knowledge Discovery (FSKD '08), Vol. 1, Pp. 52 – 56, 2008.

[26] Jain Pei, Jiawei Han, Behzad Mortazavi_asl and Hua Zhu, Mining Access Patterns Efficiently from Web Logs, Pacific-AsiaConference on Knowledge Discovery and Data Mining(PAKDD'00), Kyoto, Japan, 2000, 396-407.

[27] Heng Tao Shen, Beng Chin Ooi and Kian_Lee Tan, Giving meanings to WWW, ACM SIGM Multimedia, L.A,2000, 39-47.

[28] Jia-ching Ying, Vincent S. Tseng, Philip S. Yu IEEE International Conference on Data Mining workshops IEEE Computer Society, 2009.

[29] Jing Wang, Ying Liu, Lin Zhou, Yong Shi, and Xingquan Zhu, Pushing frequency constraint to utility Mining Model, ICCS Springer-Verlag Berlin Heidlberg, LNCS 4489, 2007, 685-692

[30] C. Ridings and M. Shishigin, Pagerank Uncovered, Technical report, 2002.

[38] Neelam Duhan, A.K. Sharma and Komal Kumar Bhatia, Page Ranking Algorithms: A Survey, IEEE International Advance Computing Conference, Patiala, 2009, 1530-1537.

[31] Geeta.R.B, Shashikumar G. Totad & Prasad Reddy PVGD, Topological Frequency Utility Mining Model Springer International Conference, SocPros 12, 2011, 505-508.

[32] Chu-Hui Lee, Yu-lung Lo, Yu-Hsiang Fu, "A novel prediction model based on hierarchical characteristic of web site", Expert Systems with Applications 38, 2011.

[33] V. Sujatha, Punithavalli, "Improved User Navigation Pattern Prediction Technique From Web Log Data", Procedia Engineering 30, 2012.

[34] A. Anitha, "A New Web Usage Mining Approach for Next Page Access Prediction", International Journal of Computer Applications, Volume 8– No.11, October 2010

[35] Mehrdad Jalali, Norwati Mustapha, Md. Nasir Sulaiman, Ali Mamat, "WebPUM: A Web-based recommendation system to predict user future movements" Expert Systems with Applications 37, 2010.

[36] TrilokNathPandey, Ranjita Kumari Dash, Alaka Nanda Tripathy ,Barnali Sahu, "Merging Data Mining Techniques for Web Page Access Prediction: Integrating Markov Model with Clustering", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 6, No 1, November 2012.

