International Journal for Modern Trends in Science and Technology
Volume: 03, Issue No: 02, February 2017
ISSN: 2455-3778
http://www.ijmtst.com

A Survey on Preprocessing of Web-Log Data in Web Usage Mining

A V Srinivas

Assistant Professor, Department of IT, Vardhaman College of Engineering, Hyderabad, Telangana, India.

To Cite this Article
A V Srinivas, "A Survey on Preprocessing of Web-Log Data in Web Usage Mining", *International Journal for Modern Trends in Science and Technology*, Vol. 03, Issue 02, 2017, pp. 35-41.

ABSTRACT

Web mining is to determine and extract useful information. In the internet age web applications are increasing at enormous speed and the web users are increasing at exponential speed. As number of users grows, web site publishers are having increasing their information for attracting and satisfying users. It is possible to trace the users' essence and interactions with web applications through web server log file and Web log file contains only (.txt) file. The data stored in the web log file consist of large amount of eroded, incomplete, and unnecessary information. Because of large amount of irrelevant data's available in the web log file, an original log file cannot be directly used in the web usage mining. So preprocessing technique is applied to improve the quality and efficiency of a web log file. Different techniques are applied in preprocessing that is data cleaning, data fusion, data integration. In this paper we will survey different preprocessing technique to identify the issues in web log file and to improve web usage mining preprocessing for pattern mining and analysis.

KEYWORDS: *data cleaning, data fusion, data integration, Preprocessing technique, Web usage mining, web log file.*

I. INTRODUCTION

During the past few years the World Wide Web has become the biggest and most popular way of communication and information dissemination. It serves as a platform for exchanging various kinds of information. The volume of information available on the internet is increasing rapidly with the explosive growth of the World Wide Web and the advent of e-Commerce. While users are provided with more information and service options, it has become more difficult for them to find the "right" or "interesting" information, the problem commonly known as information overload. Web mining is the application of data mining techniques to extract knowledge from Web data including Web documents, hyperlinks between documents, usage logs of web sites, etc. A common taxonomy of web mining defines three main research lines: content mining, structure mining and usage mining.

Web content mining is the process to discover useful information from the content of a web page. Basically, the Web content consists of several types of data such as textual, image, audio, video, metadata as well as hyperlinks.

Web Structure Mining is the process of inferring knowledge from the World Wide Web organization and links between references and referents in the Web. The structure of a typical web graph consists of web pages as nodes and hyperlinks as edges connecting related pages. Web Structure mining is the process of using graph theory to analyze the node and connection structure of a web site.

Web usage mining also known as web log mining is the application of data mining techniques on large web log repositories to discover useful

knowledge about user's behavioral patterns and website usage statistics that can be used for various website design tasks. The main source of data for web usage mining consists of textual logs collected by numerous web servers all around the world. There are four stages in web usage mining.

Data Collection*:* users log data is collected from various sources like serverside, client side, proxy servers and so on.

Data Preprocessing:Performs a series of processing of web log file covering data cleaning, user identification, session identification, path completion and transaction identification.

Pattern discovery*:* Application of various data mining techniques to processed data like statistical analysis, association, clustering, pattern matching and so on.

Pattern analysis*:* once patterns were discovered from web logs, uninteresting rules are filtered out. Analysis is done using knowledge query mechanism such as SQL or data cubes to perform OLAP operations.

Typically three main data sources are used to collect log data for web usage mining. Those are Server log, Proxy server log, Client/ Browser log.

Server Log*:* When an internet user request a particular page on web, an entry is logged into a special file called server log file. This file is not accessible by general internet user, only administrative person or server owners can access these files [1]. Server logs are considered as a richest and reliable source of information to predict user's behavior but it lacks with many quality factors such as completeness and privacy issues.

Proxy Server Log:A Proxy server is a server which acts as an intermediary between user's requests to other web servers. They are generally used for caching services to improve navigation speed, administrative control and security. Collecting proxy level usage data is similar as collecting server level data.

Client/Browser Log: Web log data can also be collected from client machine by integrating java applets to the website, writing java scripts or even modified browsers. Client side logs are useful to tackle problems related with server logs like web page caching, session reconstruction [2,3].

II. Data Preprocessing

The information available in the web is heterogeneous and unstructured. Therefore, the preprocessing phase is a prerequisite for discovering patterns. The goal of preprocessing is to transform the raw click stream data into a set of user profiles [4]. Data preprocessing presents a number of unique challenges which led to a variety of algorithms and heuristic techniques for preprocessing tasks such as merging and cleaning, user and session identification etc [5]. Various research works are carried in this preprocessing area for grouping sessions and transactions, which is used to discover user behavior patterns.

A. Data Cleaning:

Data Cleaning is a process of removing irrelevant items such as jpeg, gif files or sound files and references due to spider navigations. Improved data quality improves the analysis on it. The Http protocol requires a separate connection for every request from the web server. If a user request to view a particular page along with server log entries graphics and scripts are download in addition to the HTML file. An exception case is Art gallery site where images are more important. Check the Status codes in log entries for successful codes. The status code less than 200 and greater than 299 were removed.

B. User Identification:

Identification of individual users who access a web site is an important step in web usage mining. Various methods are to be followed for identification of users. The simplest method is to assign different user id to different IP address. But in Proxy servers many users are sharing the same address and same user uses many browsers. An Extended Log Format overcomes this problem by referrer information, and a user agent. If the IP address of a user is same as previous entry and user agent is different then the user is assumed as a new user. If both IP address and user agent are same then referrer URL and site topology is checked. If the requested page is not directly reachable from any of the pages visited by the user, then the user is identified as a new user in the same address [6]. Caching problem can be rectified by assigning a short expiration time to HTML pages enforcing the browser to retrieve every page from the server [7].

C. Session Identification:

A user session can be defined as a set of pages visited by the same user within the duration of one particular visit to a web-site. A user may have a single or multiple sessions during a period. Once a user was identified, the click stream of each user is portioned into logical clusters. The method of portioning into sessions is called as Sessionization or Session Reconstruction. A transaction is defined

as a subset of user session having homogenous pages.

D. Path Completion:

There are chances of missing pages after constructing transactions due to proxy servers and caching problems [8] [9]. So missing pages are added as follows: The page request is checked whether it is directly linked to the last page or not. If there is no link with last page check the recent history. If the log record is available in recent history then it is clear that "back" button is used for caching until the page has been reached. If the referrer log is not clear, the site topology can be used for the same effect. If many pages are linked to the requested page, the closest page is the source of new request and so that page is added to the session.

III. Literature Survey

Commonly used sub-steps of preprocessing are Data cleaning, User identification, Session identification and Path completion. Different researchers have introduced various preprocessing techniques to improve efficiency and scalability of pattern discovery techniques. Some of them are discussed below:

Cooley et al. [10] have proposed methods for data cleaning, user identification, session identification and transaction identification. Although their methods are good enough but some heuristics are not appropriate for complex web sites.

Prabarskaite [11] proposed a better cleaning methodology. According to him standard cleaning methodology is not appropriate for frame pages containing websites. He applied two approaches: advanced cleaning to improve web log mining and filtering to remove irrelevant links. In this preprocessing process author did not perform any other steps of preprocessing like user identification session identification etc.

Tanasa et al. [12] divides preprocessing process in four steps: Data fusion, Data cleaning, Data structuration and Data summarization. In Data fusion author joined multiple log files from different web servers and also from site maps into a single log files. After that they anonymized log file by encrypting host name. Further Data cleaning is performed by removing requests for non-analyzed resource such as multimedia files (images, audio, video etc.) and robot's generated requests In Data structuration part author have completed user identification by authentication data or IP address, Session identification by host and agent, Page view identification by site map etc. At last Data summarization step includes pattern analysis part by using data generalization and aggregation. They did not considered unsuccessful request in data cleaning phase which is also required to remove to get rid of unnecessary calculations in later phases of web log mining processes.

Robert et al. [13] introduced a new concept called integer programming for better session identification .This method generates session simultaneously and produced session better match to an empirical distribution.

Yen li et al. [14] proposed an approach for path completion by combining Maximal forward reference length and Reference length algorithm. First Maximal forward reference is used to find the sequence of page in user access path and it is also used to identify the page, and finally Reference length algorithm is used to find whether the page is informative page or auxiliary page. Lastly by using referrer field complete path has been built.

Xiang–ying li [15] has proposed an algorithm named CSIA (Client and Session Identification algorithm) for identification of user and sessions. This algorithm includes comprehensive approach by combining IP address, topology, browser version and referrer page to identify unique user with better accuracy and efficiency. He proposed his algorithm in JAVA language framework as it is good for space utilization. However this algorithm is suffering with decrease in operating rate due to consideration of many factors for identifying user.

Sanjay Bapu Thakare et al.[16] described the effective and complete preprocessing of access stream before actual mining process can be implemented. They suggested Improvised merging algorithm for data preprocessing method for margining process. They did implementation of field's extraction in core Java. They also suggested improvised TransLog algorithm for convert log file in database. They did not discuss about user identification and more processes of web usage mining. For implementation they used IIS log file format.

Amit Dipchandji Kasliwal et al. [17] proposed a web usage mining methods using well known tool RapidMiner for predicting access behavior. They were taking a log file from KDD repository and performed web usage mining techniques. Using RapidMiner tool they did data preprocessing methods and obtaining ARFF file and applying association rule to ARFF file using MATLAB identify specific result of web pages visits.

Navin Tyagi et al. [18] surveyed about the data preprocessing activities like data cleaning, data reduction and related algorithms. They presented algorithms for data cleaning and data reduction based on CERN (Common Log Format) log file format. About further procedures of preprocessing did not mention.

K. R . Suneetha et al. [19] focused on data preprocessing techniques including web log structure, data cleaning and user identification. They used data from NASA web server log files. To improve efficiency of log files they remove unnecessary data from web log files and analyze process including generating various reports. They also did analysis on most system error found during visit of website. They did not apply any data mining algorithms for pattern discovery.

Jaideep Srivastava et al. [20] provides a detailed taxonomy of the work in area of web mining, including research area as well as commercial offerings. They provide up-to-date survey of the existing work is also provided for Web Usage Mining Research project and products. They provides idea of web usage mining applications likes personalization, system improvements, site modification, Business Intelligence and user characterization. In added in work they provides overview of WebSHIFT system which is designed to perform Web usage Mining from server logs in the extended NSCA format.

V. Chitraa et al. [21] proposed a new technique for identifying sessions for extraction of user patterns. Their experimental results show that the proposed Session Identification technique is an effective one to construct sessions accurately. In their proposed method a matrix is constructed from which sessions are identified using MATLAB tool. They proposed session construction algorithm based on browsing time. They mainly focus on preprocessing web data using data cleaning, user identification and session identification.

Sheetal A. Raiyani et. al. [22] Introduced proposed Technique DUI (Distinct User Identification) based on IP address, Agent, Referred pages on desired session time. They discussed all methods of preprocessing including web log format. They used R K University's library web log of 8 months from Jan1, 2012 to Aug 3, 2012 and implemented preprocessing techniques. In result they got different distinct user's number on their month wise data.

Vellingiri J. et al.[23] focuses on providing techniques for better data cleaning and transaction identification from the web log. They used data preprocessing methods including data cleaning by remove unnecessary data, robot cleaning; user identification using reference length, where reference length is the time taken by the user to view a particular page; session identification, path completion and transaction identification using reference length. They focused on two algorithms one is Maximal Forward References (MFR) and Reference Length (RL). Using these two algorithm author helps to determine only the relevant logs that the user is interested in.

G. T Raju et al.[5] proposed a complete preprocessing methodology that allows the analyst to transform any collection of web server log files into structured collection of tables in relational database model. They compared preprocessing techniques which used by researchers including data source, data cleaning and data formatting and structuring. They showed experiment results likes reducing size of log file for preprocessing, day wise unique visitors, user session identifications.

P. Nithya et al.[24] proposed a novel pre-processing technique by removing local and global noise and web robots. They implemented data cleaning phase will helps in determining only the relevant logs that the user is interested in. Anonymous Microsoft Web Dataset and MSNBC.com. They did not mention other preprocessing methods like user identification, session identification and path completion.

Vellingiri J. et al.[25]provided three phases of web usage mining for user navigation discovery including preprocessing phase, Identify user's behavior and classifications of user behaviors. In the preprocessing phase, the data cleaning process includes removal of graphics, video, status code and robots cleaning. In the second phase, design a set of clusters using Weighted Fuzzy-Possibilistic C-Means (WFPCM), which consists of "similar" data items based on the user behavior and navigation patterns for the use of pattern discovery. In the third phase, classification of the user behavior is carried out for the purpose of analyzing the user behavior using Adaptive Neuro- Fuzzy Inference System with Subtractive Algorithm (ANFIS-SA).

Ashwin Riyani et al.[26]focused on a complete preprocessing style having data cleaning, user and session Identification activities to improve the quality of data. They introduced proposed technique algorithm) DUI (Distinct User Identification) based on IP address,Agent and Session time, Referred pages on desired session time. They did not use pattern discovery and pattern analysis methods.

Renáta Iváncsy et al.[27]provided novel approach that uses a complex cookie-based method to identify web users. They developed an implementation called Web Activity Tracking (WAT) system that aims at a more precise distinction of web users based on log data. Using WAT they presented different useful result.

Arvind Dangi et al.[28] proposed a new method for web data preprocessing in which it has three phases. In the first phase some websites are selected and by different locations access these website & by applying the (java) tools & methods then find out the IP address of that websites, session usage time & navigations, in the final phase combine them as framework which may be helps to investigate the web user usage behavior.

S. Prince Mary et al.[29] describes the preprocessing methods and steps involved in retrieving the required information effectively. They included data collection, data cleaning by local and global noise removal & graphics records, HTTP failure status etc, User and session identification with path completion.

Shaily Langhnoja et al.[30] gives detailed description of how preprocessing is done on web log file and after that it is sent to next stages of web usage mining. They Used algorithms of data cleaning, user and session identification on web log files and showed results.

Zidrina Pabarskaite [11] presented two new techniques for enhancing web log mining. First is novel framework for performing advanced web log data cleaning and second is data mining is visualization.

C. E. Dinuca [31] propose a new method for identifying sessions based on average time of visiting web pages based on the use of fixed values cause errors in identifying sessions. They implemented in Java programming language by using NetBeans IDE and used two algorithms to identify sessions including 30 minutes to indicate end of session and average time spent on page by users. They showed result and conclude that complexity of classify algorithm is not modified by new approach.

IV. Conclusion

In last few decades web has become an informational hub for users. Thus analysis of user's behavior is becoming more and more important for e-commerce companies to provide better services to customers and visitors. Web usage mining is a field of study where user's activity is analyzed and processed to generate useful patterns. Due to irrelevant data in log file, preprocessing is considered as an essential step in web usage mining. In this paper, different steps of preprocessing: Data cleaning, User identification, Session identification, and Path completion have been discussed. Web usage mining depicts various challenging problems for preprocessing of log data. High dimensionality and large volume of data results in high computational complexity of mining process. So there is need to compress data without losing essential information regarding user's behavior. On basis of review,, it can be seen that there are many common data preprocessing techniques applied in various types of log files. Some authors used common techniques like remove graphical records, HTTP failure status record etc. But three or four authors included robot cleaning in data cleaning preprocessing which helps while log files are collected from proxy server or having proxy server between server and client. To identify user, researchers used different methods based on relevant data with predicting IP address relationship, based on cookie, based on reference length, etc. Researcher implemented different algorithms for identifying session based on 30 minutes expiration time or average time spent by users, etc. Further, it is found that there is no algorithm focusing on multiple clients accessing website through the same IP address. It can be considered an area of further research.

.

References

[1] Zidrina Pabarskaite, Aistis Raudys (2007), A process of knowledge discovery from web usage data: Systemization and critical review in Journal of Intelligent Information System, Springer Vol.28 Issue.1 Page(s): 79-104.

[2] C. Shahabi, F. Banaei-Kashani (2002), A Framework for Efficient and Anonymous Web Usage Mining Based on Client-Side Tracking in WEBKDD Third International Workshop on Mining Web Log Data, Page(s): 113-144.

[3] D. Pierrakos, G. Paliouras, C. Papatheodorou, and C. D. Spyropoulos (2003), Web usage mining as a tool for personalization: A survey in User Modeling and User Adapted Interaction journal, Vol. 13 Issues. 4 Page(s): 311-372.

[4] B. Naresh Kumar Reddy, M.H.Vasantha, Y.B.Nithin Kumar and Dheeraj Sharma, "Communication Energy Constrained Spare Core on NoC", 6th International Conference on Computing, Communication and Networking Technologies (ICCCNT), PP. 1-4, 2015.

[5] B. Naresh Kumar Reddy, M.H.Vasantha, Y.B.Nithin Kumar and Dheeraj Sharma, "A Fine Grained

Position for Modular Core on NoC IEEE International Conference on Computer, Communication and Control, PP. 1-4, 2015.

[6] Robert.Cooley,Bamshed Mobasher and Jaideep Srinivastava,"Data Preparation for Mining World Wide Web Browsing Patterns ,", journal of knowledge and Information Systems,1999.

[7] Cyrus Shahabi, Amir M.Zarkessh, Jafar Abidi and Vishal Shah "Knowledge discovery from users Web page navigation, ", In.Workshop on Research Issues in Data Engineering, Birmingham, England,1997.

[8] Yan Li, Boqin FENG and Qinjiao MAO, "Research on Path Completion Technique in Web Usage Mining,,", International Symposium on Computer Science and Computational Technology, IEEE,2008.

[9] Yan Li and Boqin FENG "The Construction of Transactions for Web Usage Mining,,", International Conference on Computational Intelligence and Natural Computing, IEEE,2009.

[10] R. Cooley, B. Mobasher, J. Srivastav (1999), Data preparation for mining world wide web browsing pattern in Journal of Knowledge and Data Engineering Workshop, IEEE, Vol.1 Page(s): 5-32.

[11] B. Naresh Kumar Reddy, M.H.Vasantha and Y.B.Nithin Kumar, "A Gracefully Degrading and Energy-Efficient Fault Tolerant NoC Using Spare core", 2016 IEEE Computer Society Annual Symposium on VLSI, pp. 146-151, 2016.

[12] D. Tanasa, B. Trousse (2004), Advanced Data Preprocessing for Intersites Web Usage Mining in IEEE Intelligent Systems, Vol. 19 Issues. 2 Page(s): 59-65.

[13] R. F. Dell (2008),Web user session reconstruction using integer programming in International Conference on Web Intelligence and Intelligent Agent Technology, IEEE/ACM/WIC, Vol. 1 Page(s): 385-388.

[14] Yan LI (2008), Research on path completion technique in web usage mining in International Symposium on Computer Science and Computational Technology, IEEE, Vol. 1 Page(s): 554-559.

[15] Xiang-ying Li (2013), Data Preprocessing in Web Usage Mining in 19th International Conference on Industrial Engineering and Engineering Management Page(s): 257-266.

[16] Sanjay Bapu Thakare, Prof. Sangram Z Gawali, "A Effective and Complete Preprocessing for Web Usage Mining", International Journalon Computer Science and Engineering, Vol. 02, No. 03, pp. 848-851,2010,

[17] Amit Dipchandji Kasliwal, Dr. Girish S. Katkar, " Web Usage mining for predicting User Access Behavior", International Journal of Computer Science and Information Technology, Vol. 6 (1), 2015, 201-204

[18] Navin Kumar Tyagi, A.K.Solanki, Sanjay Tyagi, " An Algorithmicapproach to data preprocessing in Web Usage Mining", InternationalJournal of Information Technology and Knowledge Management, Vol.2, No. 2, pp. 279-283,Dec-2010

[19] K.R. Suneetha , Dr. R. Krishnamoorthi, "Indentifying User Behavior byanalyzing Web Server Access Log File", International Journal ofComputer Science and Network Security, Vol. 9, No. 4, April 2009

[20] Jaideep Srivastava, Robert Cooley, Mukund Despande, Pang-Ning Tan,"Web Usage Mining: Discovery and Applications of Usage Patternsfrom Web Data", SIGKDD Explorations, Vol. 1, Issue 2 Jan 2000

[21] V.Chitraa, Dr. Antony Selvadoss Thanamani, "Web Log Data Cleaning for Enhancing Mining Process", International Journal of Communication and Computer Technologies", Vol. 01, No. 11, Issue 03, December 2012.

[22] B. Naresh Kumar Reddy, et al., An Efficient Data Transmission by using Modern USB Flash Drive", International Journal of Electrical and Computer Engineering, Vol. 4, Issue 5, 2014.

[23] Vellingiri J. and S. Chenthur Pandian, " A Novel Technique for Web Log mining with Better Data Cleaning and Transaction Identification ",Journal of Computer Science 7(5): 683-689, ISSN: 1549-3636, 2011

[24] P. Nithya and Dr. P. Sumathi, "Novel Pre-Processing Technique for Web Log Mining by Removing Global Noise, Cookies and Web Robots", International Journal of Computer Applications, Vol. 53, No.17, September-2012

[25] Vellingiri J., S. Kaliraj, S. Satheeshkumar and T. Parhtiban, " A Novel Approach for User Navigation Pattern Discovery and Analysis for Web Usage Mining", Journal of Computer Science 11(2); 372-382, 2015

[26] Ashwin G. Raiyani, Sheetal S. Pandya, "Discovering user identification mining techniques for preprocessed web log data", Journal ofInformation, Knowledge and Research in Computer Engineering, ISSN: 0975-6760, Vol. 2, Issue. 2, Pages 477-482, OCT-2013

[27] Renata I., Sandor J., "Analysis of Web User Identification Methods",World Academy of Science, Engineering and Technology, 2007

[28] Arvindkumar Dangi, Sunita Sangwan, " A new approach for user identification in web usage mining preprocessing", IOSR Journal of Computer Engineering, e-ISSN: 2278-0661, p-ISSN: 2287-8727, Vol. 11, Issue. 3, (May-June2013), Pages 57-61

[29] Vijaya Sree Boddu and et. al., "Low power and area efficient N-bit parallel processor on a chip", 13th International IEEE India Conference INDICON 2016, pp. 1-4, 2016.

[30] Shaily Langhnoja, Mehul Barot, Darshak Mehta, " Pre-processing: Procedure on Web Log File for Web Usage Mining", InternationalJournal of Emerging Technology and Advanced Engineering, ISSN:

2250-2459, ISO 9001:2008 Certified Journal, Vol. 2, Issue. 12,December 2012

[31] C.E. Dinuca, D. Ciobanu, " Improving the session identification using the mean time", International Journal of Mathematical Models andMethods in Applied Sciences, Vol. 6, Issue 2, 2012