Volume 6, Issue 6, June 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Techniques of Web Usages Mining

Priyanka Rathod, Prof. R. R. Keole
CSIT-SGBAU, Amravati, Maharashtra,
India

Abstract- The World Wide Web (WWW) has influenced a lot to both users (visitors) as well as the web site owners. Enormous growth of World Wide Web increases the complexity for users to browse effectively. To increase the performance of web sites better web site design, web server activities are changed as per users' interests. To achieve this they have to analyze user access pattern which are captured in the form of log files. Web usage mining is a process of analyzing interaction of user with different web application. In this paper, we provide detailed survey of work done so far on data collection and pre-processing stage of web usage mining.

Keywords: Data mining, Web usage mining, Web log mining, Preprocessing

I. INTRODUCTION

In Current era, internet is playing such a vital role in our everyday life that it is very difficult to survive without it. World Wide Web (WWW) has been proving to be tremendous amount of data and also data on WWW is growing exponentially in terms of both their size and its usage with respect to time. In contrast to the standard data mining methods web data mining methods need to deal with heterogeneous, semi structured or unstructured data [1]. In Web Data Mining various core or applied data mining techniques are applied to obtain some interesting knowledge out of data available on WWW. Also the resources (web pages) on WWW undergo frequent updation in terms of their content, structure, with respect to time. Web data mining can be categorized based on the interest and/or final objective of what kind of knowledge to mine from web data [2]. **1) Web Content Mining:** refers to discovery of useful information or knowledge from web page contents i.e. text or it could be multimedia data like image, audio, video etc. **2) Web Structure Mining** aims at analyzing, discovering andmodeling link structure of web pages and/or web site to generate structural summary on which various techniques are applied and outcomes of these techniques can be utilized to recreate, redesign the web site which ultimately improves structural quality of web site [3]. **3) Web Usage Mining** deals with understanding of user behavior, while interacting with web site, by using various log files to extract knowledge from them.

II. DATA COLLECTION

There are three main sources to get the row log data, which are namely 1) Client Log File 2) Proxy Log File 3) Web Server Log File [6].

A. Web Server Log File:

The most significant and frequently used source for web usage mining is web server log data. This web log data is generated automatically by web server when it services user request, which contains all information about visitor's activity [2]. The common server log file types are access log, agent log, error log and referrer log [6] Table-1 summarizes each. Depending on web server, web log file data varies on number, type of attributes, and format of log file [7]. W3C maintains standard log file format however custom log file format can be configured. Many varied format are available like 1.Common log format, 2.Extended common log format, 3. Centralized log format, 4.NCSA common log format, 5.ODBC logging, 6.Centralized binary logging. [8]. among all common or extended file format are mainly implemented by web server.

```
#Software: Microsoft Internet Information Services 6.0
#Version: 1.0
#Date: 2008-09-16 16:17:18
#Fields: date time s-sitename s-ip cs-method cs-uri-stem cs-uri-query s-port cs-username c-ip cs(User-Agent) sc-status sc-substatus sc-win32-status
2008-09-16 16:17:18 W3SVC1 127.0.0.1 GET /BoardEzLog/Service1.asmx WSDL 80 - 127.0.0.1 Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.2;+SV1;+.NET+CLR+1.1.4322) 200 0 0
```

Fig. 1: W3C Extended Common Log Format (ECFL) file [20**].**

W3C Extended Log File Format (Figure-1) is very valuable in web usage mining as it can be customized. It contains some additional attribute then CLF [7, 20]. These are i) REFERER_URL defines the URL where visitor came from. ii) HTTP_USER_AGENT reflects visitor's browser version, iii) HTTP_COOKIE is a persistent token, used to identify user

uniquely, which is sent to visitor. Web Server may use caching for efficiency purpose. So if request comes from user for a particular page and if this page is there in its cache it will be delivered to the user without making entry into the web server log file.

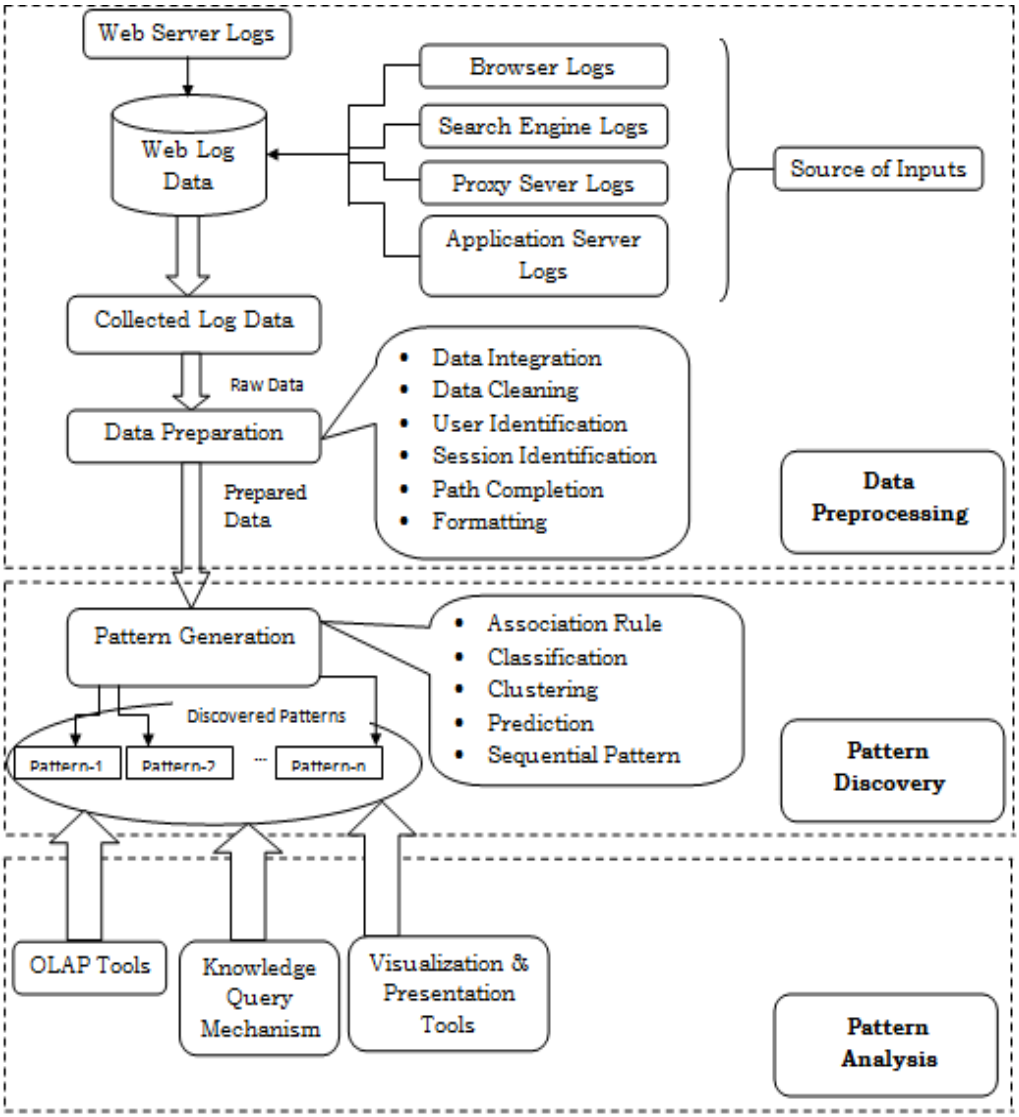

Fig. 2 Web Usage Mining Process

B. Client Side Log File:

Refer to recording of activities, events that happens within the premises of client machine. Like mouse wheel rotation, scrolling within a particular page, mouse clicks, content selection [9]. In some case it is advantageous, as it eliminates necessity of session identification, caching [11]. This can be recorded by number of ways:

1) By integrating java applet with web site: which records of the activity of users. But for that java plug in need to be installed on each client side browser. Also user may experience delay in page loading time, when applet is loaded for the first time [11].

2) By writing Java Scripts: in almost each pages of web site that will record this interaction of user with web page and report it to server when transaction is complete. This approach requires each page to be re-created, re-designed which could be time consuming, cumbersome even in some case not technically feasible because of the limitations of web hosting and allied server side software/ hardware components

3) By developing a browser plug-in: which need to be installed only once which can record this interaction and will send the record at finite interval of time or just before when user is about to close the connection with website or when user is quitting from browser. This can be done without changing the underlying design, architecture or technology of web site. However user's collaboration is required and compatible plugins needs to be developed per browser type. [9, 10] demonstrated how client side public or private data like content of my documents, calendar, browser history, favorites, bookmarks can be used for WUM application like User Profiling and Content-based Recommendation. [10] Suggested a system, which does recommendation, consisting of three tiers (layer). Layer-1 is row information collection agent, which collects data from client machine. Layer-2, a logic layer uses this data to create Dynamic User Profile (DUP), Layer-3 is responsible for presentation and customized UI. [9] Suggested to build such a dynamic profile from various hardware level events like keyboard, mouse etc.

C. Proxy Server Log File:

At many places network traffic is routed through a dedicated machine known as a proxy server, all the request and response are serviced through this proxy server. Study of this proxy server log files, whose format is same as of web log file may reveal the actual HTTP requests coming from multiple clients to multiple web servers and characterizes, reveals the browsing behavior for a group of anonymous users sharing a common proxy server [11]. Some web sites use n-tier architecture to have reliable, efficient and secure web applications. Log data that are gathered at application server while servicing the users request can also be used for web usage mining. They peculiarly show how user requests are serviced and may assist in identifying and understanding the internal calls-page access resulted to fulfill a single request. Entire process of web usage mining can be logically divided into three significant and co-related steps as shown in Figure-2, which is Data Preprocessing (Data Preparation), Pattern Discovery (Knowledge Discovery).Pattern Analysis (Knowledge Analysis & Presentation).

III. DATA PREPROCESSING

Due to diversity of sources individual or obtained combined log file, which contains row log data is unformatted, may contain noise, impurities and directly on it [5]. So a row log data undergoes a complex process, consisting of series of steps/stage called Data Preprocessing. It removes such impurities and /or converts data into format on which data mining techniques can be applied [7]. It aims to build and provide a reliable, robust structural framework on which success of later stage relies, which is application of various data mining techniques (Pattern Discovery) [12]. Data Preparation is the most complicated and time consuming task. About 80 percentages [13] of time is given on this process to strengthen quality of data because as qualitative the data is better the results. For this data preparation task which mainly includes various sub-task namely data cleaning, user identification, session identification, path completion and transaction identification [12]. Plenty of algorithms, heuristic techniques are developed and suggested for this, using which a robust, reliable and integrated data source can be created and later on various data mining technique can be applied on them efficiently. Depending on what to mine any above listed sub task can be repeated or eliminated at all. Here we provide an in depth review and work done on data Preprocessing methods.

A. Data Cleaning & Feature selection:

It is a process of identifying, selecting and removing of unnecessary or irrelevant fields and/or rows form row log data. Web log file contains so many attributes (fields) only necessary fields are selected rest of them are dropped. Firstly entries for access of JPEG, GIF file, Java Scripts, other audio/video files need to be removed as they are executed or downloaded not on basis of user's request and hence might be redundantly recorded in log files. Secondly if user requests a page or resource which is not available on web server, those entries are marked with different status code (error), which also needs to be discarded. Thirdly the entries occurred from the crawlers or spiders also need to be eliminated because they do not reflect the way human visitor navigate the site. Many crawlers declare themselves as an agent and hence can be detected easily by simple string matching. [14] Employs various heuristic based on which non-human behavior can be detected. [7] Suggest that records which are too rear or too frequent will not lead to constitute any meaningful or important knowledge from it. For example records pertaining with access to index.html or home.html are not of much interest and hence can be dropped. Table-2 summarizes data cleaning

IV. PATTERN DISCOVERY

It is the ultimate stage where some useful knowledge will be derived by applying various statistical and/or data mining techniques at hand from various research areas like data mining, machine learning, statistical method and pattern recognition. Frequently used techniques are classification, clustering, association rule, sequential pattern etc [4, 5, 24]. *Clustering* aims to build clusters and categorize users in to groups (clusters) who demonstrated similar browsing behavior, also known as user clustering [7]. Page clustering techniques indentifies group of pages which are conceptually related. It can be done by measuring similarities between two entities. Some commonly used techniques are Euclidian Distance, SPO, and Fuzzy C-Mean etc [19, 21]. Clustering forms base for the web personalization, adoption to an individual user' need. Based on clustering user demographic behavior, market segmentation for an E-Commerce site, recommendation can be planed and delivered in a personalized way [11]. *Classification* is considered as supervised learning. It is an automated process of assigning a class label or mapping a user based on browsing history or on the basis of some other attribute with one of existing class. It can be done by various inductive learning algorithm like decision tree classifier, naïve Bayesian classifiers, support vector machine It forms the bases for WUM application like profile building. Later on based on classified user profile efficient personalization, recommendation can be made [7, 9, 10]. *Association Rules* are able to discover related item occurring together in same transaction, and is used to find interdependency, co-relation among the pages. Such number of rules generated could be very large so two measures support and confidence is employed, which determines importance and quality of rules [7, 11]. A-Priori and its many versions are developed to mine association rule. *Sequential patterns* (rule) are formed when we attach a time domain with some other attribute of interest. The problem of mining sequential patterns is to find the maximal frequent sequences among all sequences that have a certain user specified minimum support [7]. Using this web marketer can better match advertisement with targeted user groups [11].

V. PATTERN ANALYSIS

Result of pattern discovery phase might not be in the form, suitable for interpretation or to derive conclusion out of it. It provides ways to compare the results and to extract interesting rule or pattern from output of previous step [25]. So

various visualization and presentation tools are used which represent data in 2D, 3D pictorial representation. This tool provides interactive way of representing, comparing, characterizing result in terms of charts, graphs, tables, vein diagram and so many others visual presentations [25]. Many times result generated or data itself are stored in data cubes or in data ware house on which various OLAP operations such as roll-up, drill-down, slice etc. can be performed which provides multiple view of same data to analyzer in logical and hierarchical structure. Knowledge Query Mechanism such as SQL facilitates to retrieve data in a way controlled by analyzer, generally kind of statistical data in text format.

VI. CONCLUSION

Web sites are of much use for users. Web sites are built, deployed and maintained to serve with various function to user. At what extent this functions, features which were thought of is implemented can be identified, verified by careful inspection at the log data. Based on this result further corrective, measuring action can be planned, executed. To be able to achieve this knowledge is accomplished through the application of various subjective and/or objective, procedural algorithmic or heuristic processes, methods or techniques.

VII. FUTURE WORK

Web log data pre-processing is very important and crucial task in entire process. This phase can be strengthened by choosing and neatly applying various heuristic techniques. Most of the systems, architecture that were implemented or proposed considers either client side or server side log data. In future a system could be build that considers and exploit the usefulness of both client side and server side log data, to produce result that are more efficient and batter match with empirical observations.

ACKNOWLEDGMENT

I express my sincere thanks to **Prof P.L.Ramteke, HOD of CSE dept of HVPM, COET, Amravati** for allowing me to present this paper also I express my heartiest thanks to **Prof. R.R.keole Assist. Prof. of HVPM, COET, Amravati** for his valuable guidance while preparing this paper and guiding me time to time. Also all the friends and Staff who help me in preparing this paper.

REFERENCES

[1] Qingyu Zhang, Richard Segall "Web mining: a survey of current research, techniques and software", International Journal of Information Technology & Decision Making Vol. 7, No. 4, 2008.

[2] Kosala and Blockeel, "Web Mining Research: A Survey", SIGKDD Exploration, Newsletter of SIG on Knowledge Discovery and Data Mining, ACM, Vol.2, 2000.

[3] B. Singh, H. K. Singh, "Web Data Mining Research: A Survey", IEEE, 2010.

[4] R. Cooley, B. Mobasher, J. Srivastava, "Web mining: information and pattern discovery on World Wide web", tools with artificial intelligence, Ninth IEEE International November 1997.

[5] J. Srivasta, R.Cooley, M.Deshpande, P.Tan, "Web usage mining: discovery and applications of usage patterns from Web data", ACM SIGKDD Vol.7, No.2, Jan-2000.

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

[7] Zidrina Pabarskaite, Aistis Raudys, "A process of knowledge discovery from web log data: Systemization and critical review", Journal of Intelligent Information System, Springer, 2007.

[8] S. K. Pani et al.," Web Usage Mining: A survey on pattern extraction from web logs", International Journal of Instrumentation, Control &Automation, Vol.1, Issue 1, 2011.

[9] Jinhyuk Choi, G Lee, "New Techniques for Data Preprocessing Based on Usage Logs for Efficient Web User Profiling at Client Side", International Conference on Web Intelligence & Intelligent Agent Technology, IEEE/ACM/WIC, 2009

[10] Ting Chen et al., "Content Recommendation System based on Private Dynamic User Profile", VIth International Conference on Machine Learning and Cybernetics, IEEE, August-2007.

[11] V. Chitra, A. S .Davamani, "A survey on preprocessing methods for web usage data", International Journal of Computer Science & Information Security, Vol.7, No.3, 2010.

[12] R. Cooley, B. Mobasher, J. Srivastav, "Data preparation for mining world

[13] wide web browsing pattern", Journal of Knowledge and Data Engineering Workshop, IEEE, 1999.

[14] S. Ansari, et al., "Integrating e-commerce and data mining: Architecture and challenges", IEEE, 2001.

[15] B. Berendt, M. Spiliopoulou, "Analyzing navigation behavior in web site integrating multiple information systems", VLDB Journal, Special issues on databases and web, 2000.

[16] J. Zhang, Ali A. Ghorbani, "The Reconstruction of user session from a server log using improved time oriented heuristic", IInd Annual Confernnce on Communication Networks and Service Research, IEEE, 2004.

[17] R. F. Dell et al., "Web user session reconstruction using integer programming", International Conference on Web Intelligence and Intelligent Agent Technology, IEEE/ACM/WIC, 2008.

[18] G. Arumugam, S. Sugana, "Optimum algorithm for generation of user session sequences using server side web user logs", IEEE, 2009.

[19] M. Heydari et al., "A graph based web usage mining method considering client side data", International Conference on Electrical Engineering and Informatics, IEEE, 2009.

[20] 10 Computational Technology, IEEE, 2008.

[21] Yan LI, Bo-qin FENG, et al., "The construction of transaction for web usage mining", International Conference on Computational Intelligencen and Natural Computing, IEEE, 2009.

[22] Jose M. Domenech1 and Javier Lorenzo, "A Tool for Web Usage Mining", 8th International Conference on Intelligent Data Engineering and Automated Learning, 2007.

[23] Liu Kewen, "Analysis of Preprocessing methods for web usage mining", International Conference on measurement, Information and Control, IEEE, 2012.

