RESEARCH PAPER | Engineering | Volume - 5 | Issue - 1 | Jan Special Issue - 2015 | ISSN - 2249-555X

Web Usage Mining: A Survey on Extracting Knowledge through Web Logs

KEYWORDS: data mining, Web usage mining, WUM, Web log mining

Suchita Patel	Prof. Mehul Barot
Post Graduate student, Department of Computer Engineering LDRP-ITR, Gadhinagar, Gujarat, India.	Assistant Professor, Department of Computer Engineering LDRP-ITR, Gandhinagar, Gujarat, India.

ABSTRACT The growth of data available online and it's users are increasing exponentially and creating complexity for users to discover knowledge on the web. To improve performance of website web site designer should modify website according to users preferences and interests. To achieve this user's navigational behavior should be analyze which is captured in log files. First of all these log files are preprocessed and converted into suitable format for applying data mining techniques on it. Various DM techniques are apply on processed log files in order to get frequent access pattern of users of website. These patterns are analyzed and used extracted knowledge in website modification, system improvement, business intelligence, website personalization. In this paper , we provide detailed survey of work done so far on process of Web Usage Mining

I. INTRODUCTION

Web mining is the application of data mining techniques to discover patterns from the web. In web mining various core or applied data mining techniques are applied to obtain some interesting knowledge [1]. It can be categorized based on what kind of knowledge you want to mine from web data:1) Web Content mining refers to discovery of useful information or knowledge from the content of web page/web site such as text or multimedia data like image, audio, video etc.[1][7], 2) Web Structure mining aims at analyzing, discovering and modeling link structure of web pages and/or web site to generate structural summary[9]. 3) Web Usage mining is the process of applying data mining techniques to discover hidden, valuable and interesting usage patterns from web data in order to understand and better serve the needs of web based application[3]. WUM is high flying due to effective use in numerous web related applications [7]. Applications of Web Usage Mining are Personalization, Website modification, System improvement, Business intelligence.

In this paper we give complete overview of what is the process of web usage mining (WUM) and work done so far on WUM process. The rest of paper is organized as follows. Section II provides some information about sources of web log data. In Section III we will give you complete overview of data preprocessing on web log. Section IV describes Pattern Discovery techniques on processed log file. Section V provides information related to Pattern Analysis. Section VI gives conclusion.

II. SOURCE OF DATA FOR WUM

When user interact with website web usage logs records user activities on web site. On the basis of spatial location based collection of user interaction record, this data may be further classified into 3 different categories 1) Web Server Log File 2) Client Log File 3) Proxy Log File [3].

A. Web Server Log File

Server log file is generated automatically by web server when it services user request, which contains all information about visitor's activity [19]. The common server log file types are access log, agent log, error log and referrer log[11].

B. Client Log File

It refers 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 [12]. From client log file we can discover true behavior of visitors.

C. Proxy Log File

At many places network traffic is routed through a proxy server, all the request and responses are services through this proxy server. Study of this log file may reveal the actual HTTP request coming from multiple clients to multiple web servers and characterized, reveals the browsing behavior for a group of anonymous users sharing a common proxy server[10].

There are many web server log file formats are available like 1) common Log File format (NCSA) 2) Extended Common Log File format(W3C) and 3) IIS Log Format (Microsoft).

III. DATA PREPROCESSSING

Raw log data is unformatted, may contain noise and impurities [19]. This is the phase where data are cleaned from noise, their inconsistencies are resolved and they are integrated and consolidated in order to be used as input to the next stage of pattern discovery [2]. The objective of preprocessing is to transfer raw log files in particular format which data mining algorithms can handle easily [3]. Data preprocessing involves various tasks as follows:

A. Data Cleaning

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[1].

- 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[1].
- 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 [1].
- The entries occurred from the crawlers or spiders also need to be eliminated because they do not reflect the

Volume - 5 | Issue - 1 | Jan Special Issue - 2015 | ISSN - 2249-555X

way human visitor navigate the site[1].
- Records which are too rear or too frequent will not lead to constitute any meaningful or important knowledge from it[16].

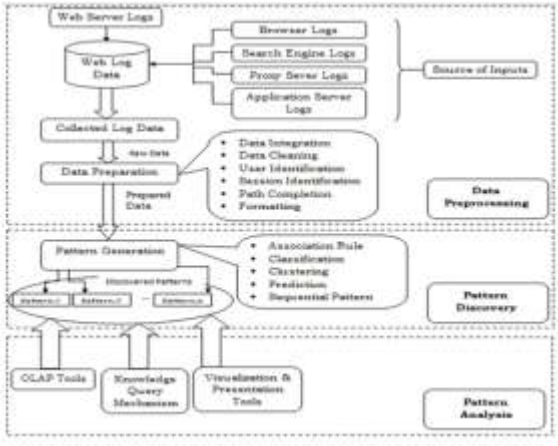

Figure 1. Web Usage Mining Process[1]

A. User Identification

User identification differentiates the log records according to users for the analysis[3]. The simplest way to identify users is to assign a different user to each different IP identified in the log file[2]. The best way to detect uniqueness of user is cookie[16] because cookies are useful for identifying the visitors of a website by storing an ID, which is generated by the web server[6]. If one is using proxy servers to route network traffic through it, web server log show a single IP address [16] but actually there are number of users who had initiated those requests which makes it difficult to identify users uniquely[1]. In such cases we can look for another way to identify users as describe below.

- Base on Agent: One possible heuristic is to look for the agent field to identify differences in OS or browser. If any one parameter is different for the records having same IP address it indicates a different user[1].
- Base on Site Topology: Also if users request a page which is not reachable from the previously visited page and if the IP address is same, it represents different user[1]. [20] Explained use of referrer attribute of W3C extended common log format to detect uniqueness.

B. Session Identification

The identification of user sessions is very important task in WUM process, as sessions encode the navigational behavior of the users and they are most important for pattern discovery[2]. A user session is a delimited set of pages visited by the same user within the duration of one particular visit to a web site [6].

Approach	Identification	Remarks
Time Oriented	Static threshold	simplicity
Navigation	Referrer field	search Site topology require Multiple time
Particle Swarm Clustering-PSO	Euclidean Distance	Good for numerical attributes
Fuzzy C-mean Clustering-FCM	Fuzzy membership function	Suitable of ill defined and overlapping boundary

Sessionization heuristics are categorized into two basic groups: 1) Time-oriented and 2) Navigation oriented. The Time-oriented heuristic applies time-out estimates to distinguish between Successive Sessions. It would be more efficient to find an appropriate time-out after analyzing the web logs, and to use different settings for each website[17].The Navigation-oriented sessionization uses either the static site structure or the implicit linkage structure captured in the referrer fields of the server logs[20] [13] explains particle swarm based clustering for web session clustering.[5] used Fuzzy C-mean clustering for session identification.

A. Path Completion

Path completion refers to the task of filling in page references that are not recorded in the web log file, due to browser and proxy server caching. Sometimes user's action does not get recorded in access log. If user clicks back word button from the browser, due to presence of cache/proxy server if local copy is present in client cache or proxy server, browser directly serves it to the user. Without making this access entry recorded in to server's web log. So this kind of missing entries preserves incomplete user path and hence requirement of detecting such missing page sequences form web logs arise, which is called path completion[1]. Such missing pages should be
mended in the log file before going for the pattern discovery[14].

To achieve this objective we need to refer referrer logs and site topology. If the referred URL of a requesting page does not exactly match with the last direct page requested, it indicates that the requested path is not complete. Further if the referred page URL is in the user's recent request history, we can assume that the user has clicked the "backward" button to visit page. But if the referred page is not in the history, it means that a new user session begins, just as we have stated above. We can mend the incomplete path using heuristics provided by referrer field and site topology[1]. [14] Proposed Reference Length algorithm (RL) and Maximal Forward Reference (MFR) in order to complete path.

IV. PATTERN DISCOVERY

In Pattern Discovery stage data mining techniques are apply on preprocessed log data in order to extract useful knowledge. Frequently used techniques for pattern discovery are:

A. Stastical Analysis such as mean, median, frequency analysis etc.

B. Clustering of users help to discover groups of users with similar navigation patterns known as user clustering [8]. Page clustering identifies group of pages which are conceptually related. It can be done by similarity measure by using PSO, Euclidean Distance, Fuzzy C- mean[13][5].

C. Classification is considered as supervised learning and it is a process of assigning a class label. In this technique knowledge is discovered by classifying users according to their navigational activities and the goal of classification is to identify the distinguishing characteristics to predefined classes, based on a set of instance e.g. users of each class[15].

D. Association Rules discovers co-relations among pages accessed together by visitor of website. Support and confidence these two measures determines quality of rules [10][16].

Volume - 5 | Issue - 1 | Jan Special Issue - 2015 | ISSN - 2249-555X

E. Sequential Patterns are formed when we attach a time domain with some other attribute of interest [1]. The problem of sequential patterns is to find the maximal frequent sequence among all sequence that have a certain user's specified minimum support[16].

V. PATTERN ANALYSIS

This is the final stage in the WUM process. In these phase discovered patterns are analyzed through OLAP tools, knowledge management: query techniques such as SQL and other visualization/ presentation tool[3]. It provides ways to compare the results and to extract interesting rule or pattern from output of previous step [4]. It enables us to do the automatic detection of patterns and by analyzing pattern we able to make predictions of new data coming from the same source.

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, wein diagram and so many others visual presentations [4].

Many times result generated or data itself are stored in data cubes or in data ware house on which various OLAP operations can be performed which provides multiple view of same data to analyzer in logical and hierarchical structure[1].

Knowledge Query Mechanism such as SQL facilitates to retrieve data in a way controlled by analyzer, generally kind of statistical data in text format[1].

VI. CONCLUSION

Web usage mining process is used to discover interesting and frequent users access patterns which are applicable to many real world problems related to website. This paper has provided detailed information about current web usage mining process and techniques to extract useful knowledge from web log files.

REFERENCE

1. Chintan R. Varnagar, Nirali N. Madhak, Trupti M. Kodinariya, Jayesh N. Rathod," Web Usage Mining: A Review on Process, Methods and Techniques",ICICES,IEEE ,Feb 2013 | 2. Mirghani. A. Eltahir,Anour F.A.Dafa-Alla ," Extracting Knowledge from Web Server Logs Using Web Usage Mining",2013 IEEE | 3. Dilip Singh Sisodia, Shrish Verma," Web Usage Pattern Analysis Through Web Logs: A Review",2012 IEEE. | 4. Liu Kewen, "Analysis of Preprocessing methods for web usage mining", International Conference on measurement, Information and Control, IEEE, 2012. | 5. Zahi Ansari, et al., "A fuzzy set theoretic approach to discover user sessions from web navigational data", IEEE, 2011. | 6. Dr.D.Suresh Babu,"Web Usage Mining: A Research Concept of Web Mining", International Journal of Computer Science and Information Technologies, 2011. | 7. Tasawar Hussain, Dr. Sohail Asghar,Simon Fong," A Hierarchical Cluster Based Preprocessing Methodology for Web Usage Mining",IEEE 2010 | 8. International Journal of Information Technology and Knowledge Management July-December 2010, Volume 2, No. 2, pp. 279-283"An Algorithmic Approach To Data Preprocessing In Web Usage Mining ". | 9. B. Singh, H. K. Singh, "Web Data Mining Research: A Survey", IEEE,2010. | 10. 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. | 11. 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. | 12. 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 | 13. Alam, S., G. Dobbie, et al., "Particle Swarm Optimization Based Clustering Of Web Usage Data", International Conference on Web Intelligence and Intelligent Agent Technology, IEEE/ACM/WIC, 2008. | 14. Yan LI, et al., "Research on path completion technique in web usage mining ", International Symposium on Computer Science and Computational Technology, IEEE, 2008. | 15. Faten Khalil," Combining Web Data Mining Techniques for Web Page Access Prediction", University of Southern Queensland, 2008. | 16. Zidrina Pabarskaite, Aistis Raudys, "A process of knowledge discovery from web log data: Systemization and critical review", Journal of Intelligent Information System, Springer, 2007. | 17. B. Liu. Web Data Mining: Exploring Hyperlinks, Contents and Usage Data. Springer,2006. | 18. 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. | 19. Kosala and Blockeel, "Web Mining Research: A Survey", SIGKDD Exploration, Newsletter of SIG on Knowledge Discovery and Data Mining, ACM, Vol.2, 2000. | 20.R. Cooley, B. Mobasher, J. Srivastav, "Data preparation for mining world wide web browsing pattern", Journal of Knowledge and Data Engineering Workshop, IEEE, 1999.