ijecs
open access www.ijecs.in

International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 5 may, 2014 Page No. 5989-5995

Web Usage Mining For extracting Users' Navigational Behavior

Divya Racha
Lecturer, Atharva College of Engineering,
University of Mumbai, India.
Email:divya.racha22@gmail.com

***Abstract*:** In this era of internet, web sites are increasing tremendously in its volume and also becoming complex. Web Usage mining is the application of different data mining techniques on the web data which tracks user's navigational behaviors using their history of records and extracts the information of user interest using patterns. The obtained results are used in different applications such as recommender systems, business intelligence and site improvement. This paper emphasizes on the idea of better understanding of user's profile and site objectives, as well as the way users will browse web pages to better serve them withlist of web pages which are relevant to him by comparing with user's historic pattern using different web usage mining techniques.This paper also discusses different applications of Web usage mining**.**

Keywords:*Web usage mining, Web Log Preprocessing,Pattern Discovery, Pattern Analysis.*

I. INTRODUCTION

The World Wide Web (WWW) had a powerful development within the past fifteen years, having ample new websites published every day. As intermediate consequence to this growth, the quantity of data on web has been increasing at a really quick pace. Traditional data analysis techniques which helped to form informative reports by knowledge discovery however proven cumbersome. This extraordinary growth of the net triggered the event of latest domains of application, the Web Usage Mining(WUM) being one among them. Web mining is a application of different data mining techniques on data stored on web in different sources.

Web can be divided in to three categories namely Web Structure Mining (WSM), Web Content Mining (WCM) and Web Usage Mining (WUM) [1]. WSM discovers the link structure of the hyperlinks at the inter-document level of web topology. WCM mainly focuses on the structure of content of document to find useful information in the content of Web pages such as text inside a Web page, data in semi-structured form such as HTML code, multimedia data such as pictures, and downloadable files. WUM attempts to discover useful knowledge from the data especially those stored in Web log files.

Web usage mining comprises of three main stages[1] : the preprocessing phase of raw data, the user pattern discovery and the pattern analysisof results. When the user visits the web pages, user leaves behind some valuable information in web log files. Through this web log information we can find out what kind of information user wanted from the web sites using web logs. WUM can model user behavior and, therefore, to helpful in forecasting their future movements.This paper focuses on describing this phase and presents an overview of WUM and also provides a survey of the different techniques of pattern extraction used for WUM.

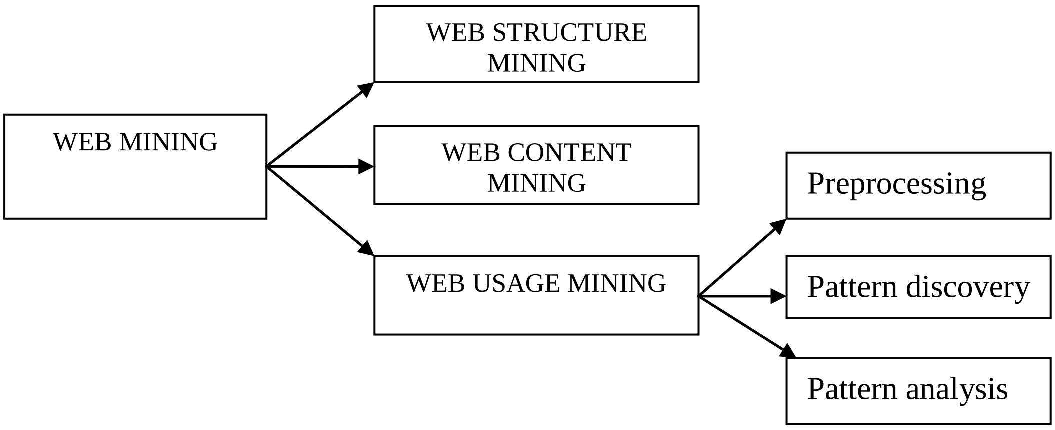

Figure 1.Taxonomy of web mining

II. Related Works

Cooley et al. [1] define Web usage mining as a three-phase process, consistingof pre-processing, pattern discovery and pattern analysis. It firstperforms intelligent cleansing and pre-processing for identifying users and server sessions. Pattern discovery is accomplished through the use of general statisticalgorithms and data mining techniques such as association rules, clusteringand classification. The preprocessed results are thenanalyzed by automatically filtering the results through a knowledge query mechanism, analytic visualization tool or the information filter to get optimized results

The WUM prediction system Analog[2]consist of two components performed online and off-line with respect to the Web server activity. The off-line component builds the knowledge base by analyzing historical data stored in server access log files, which is then used in the online component. The Online component is aimed for building the active user session and also identifies pages related to each user with list of suggestions and recommendations. The difficulty in this was to update the online and offline component which works separately.

Jianhan Zhu et al. [3] used the Markov chains to model users navigational behavior using Web structure mining. They proposed a method transition probability matrix compression for building a Markov model of a Website based on previous users behavior. The Markov model is used to make link predictions that help new users to navigate the Website.

LODAP[4] tool (log data pre-processor)which design and implemented in order to perform preprocessing oflog file. LODAP takes input log file related to website and output adatabase containing pages visited by user and identified usersessions. LODAP tool reducing size of web log file and grouping webrequest into a number of user sessions.

In order to solve some existing problems in traditional data preprocessing for web log data, an improved data preprocessing technology is used by the author ling Zheng[5]. The identification strategy based on the referred web page is adopted at the stage of user identification, that is more effective than the traditional one based on website topology. At the stage of session identification, the new strategy based on fixed priori threshold combined with session reconstruction is introduced.

The most well-known algorithm for community detection for clustering was proposed by Girvan and Newman[6]. Thismethod is historically important due to the opening a newera in the field of community detection. This method usesa new similarity measure called edge betweenness. Edgebetweenness is referred to the number of shortest pathsbetween all vertex pairs that run along that edge. Thealgorithm has a complexity O(n3) on a sparse graph.

MehradadJalali et al [7] proposed in their studies an on-line Recommendation System using LCS algorithm. For this study they used Graph Partitioning Algorithm for clustering the users and LCS algorithm for generating the list of most common set of recommendations.

One of the first research studies on Web Usage mining and semantic web is done by Bettina Berendt, Andreas Hotho and GerdStumme [8]. The study consists of two parts. The first part is related to extracting semantics from web page and the second part is to integrate the extracted semantics with web usage mining. A knowledge acquisition method called ONTEX [Ontology exploration] is extensively used in their studies

.

Author	Source of Log File	Preprocessing Technique	Algorithm Applied
D.Tanasa and B.Trousse[3]	Log Files from INRIA web sites	Data Fusion Data Cleaning Data Structuration Data Summarization	NA
Ling Zheng, HuiGui and Feng Li [4]	IIS Server Log File	Data Cleaning User Identification Session Identification Path Completion	Based on referred web page and fixed priori threshold
MehrdadJalali, Norwati Mustapha[7]	University CTI Log file	Data Cleaning User Identification Session Identification Transaction Identification	Graph partition theory, DFS
Sneha Y.S, Dr G. Mahadevan[9]	College Web Site	Data Cleaning User Identification Session Identification	Semantic web Longest common sequence
Bettina Berendt, Andreas Hotho and GerdStumme[10]	Server log	Data Filtering Session Identification	Semantic integration using ONTEX
Girvan–Newman[6]	Web site Log File	Data Preprocessing	Edge betweeness
MehradadJalali[13]	log files stored by the Web site Server	Data Transformation Data Cleaning Data Structuration Data Summarization	Longest common sequence
Sadhna K. Mishra,VineetRicharia[16]	Log file of web server	Data Preprocessing	Classification
SuleymanSalin and Pinar Senkul[14]	Log file	Data Preprocessing	Association rule and frequent pattern set

TABLE I ANALYSIS MATRIX FOR WUM

III. SOURCE OF DATA FOR WUM

Web Usage Mining use the data collected from three main sources [1]: (A) Web servers, (B) Proxy servers, and (C) Web clients.

A. Web server logs: A web server log is an important source for performing web usage mining because it explicitly records the browsing behavior of site visitors. The data in server logs reflects the access to website by multiple users. This logs can be stored in various formats are Common Log Format(CLF) or Extended Common Log Format(ECLF). Different kinds of server log are as follows:

- Website Server log: automatically created by server and maintains a history of page requests
- Packet sniffer: monitors network traffic coming to the web server and extract usage data directly from TCP/IP packets.
- Cookies: tokens generated by web server for each individual client browsers in order to automatically track the site visitors.
- CGI scripts: web server also relies on other utilities such as CGI scripts to handle data

B. Proxy server logs:A web proxy is a caching mechanism which lies between client browsers and Web servers. It reduces the load time of web pages as well as the network traffic load on both sides (i.e. server and client). Proxy server logs contain HTTP requests from multiple clients to multiple web servers. It serves as a data source to discover the usage pattern of an anonymous user groups, sharing a common proxy server.

C. The Client Side: Usage data can be tracked also on the client side by using Java Script, java applets, or even modified browsers. These techniques avoid the problems of users' session identification and the problems caused by caching. However, these approaches arise many issues concerning about the privacy laws that are quitestrict.

	1 site	Multiple sites
1 user	Javascripts or java applets	Modified browser
Multiple users	Server level	Proxy level

Table II: segments of web data

III. WEB USAGE MINING PROCESS

Web usage mining is application of data mining techniques on web data to discover user access patterns for different applications. WUM is a powerful tool to analyzing,designing and modifying the structure of website as well as it isalso useful in understanding and analyzing the site visitor'sbehavior.

Web usage mining can be divided in three different phases[4] namely,

i) Data preparation,

ii) Pattern discovery

iii) Pattern analysis

A. ***Data preparation**:* The data collected from different sources contain irrelevant and noisy data which needs to be eliminated for user and session identification. The data preparation task within the WUM process involvescleaning, structuring,filtering, summarizing data to prepare it for the pattern discovery task depending on what to mine any above listed subtask can be repeated or eliminated at all.

1. *Data cleaning module*: When requesting a web page containing additional web resources like images or script files, several implicit requests will begenerated by the Web browser. If these requests are still present when the data miningstep is performed, uninteresting patterns may be found, making the pattern analysis step more complex.By filtering of the useless data, we can reduce the web log file size to use less storage. This data cleaning module is developed to eliminate the following requests:

2. Method different from "GET":Usually entries in log files that refer to method different than "GET" are not explicit requests of the usershence eliminated.
3. Failed and corrupted requests: Erroneous files are useless for WUM and can be removed by examining the HTTP status codes.
4. Requests for multimedia objects: All log entries with an extension such as gif, JPEG, jpg, JPG, css, cgi and map in their filename should be removed since these requests do not represent the effectivebrowser activity of the user visiting the site, hence they areremoved.
5. Requests originated by Web robots:A Web robot (WR) is a software tool that periodically scans a Web site to extract its content. Since it is not human-generated requests this information is useless.
6. Reference length:Reference Length is the time taken by the user to view a particular page is calculated for pattern discovery.

1. *Data structuration module*:During this step, the requests from the raw log file are grouped by user session. A session is defined as a limited set ofresources accessible by the same user within a particular visit. Different algorithms for user and session identification can be used,

   i. User identification :For Web sites requesting registration, the user identification is the user id. For all other methods (e.g. using the IP addressor cookies)Users are identified by using these fields as follows[8].

- If two records has different IP address they are distinguished as two different users else if both IP address are same then User agent field is checked.
- If the browser and operating system information in user agent field is different in two records then they are identified as different users.

ii. Session identification:Two are based on time and one based on the navigation of users through the web pages.

•Time Oriented Heuristics**:**The following rules are used to identify a session:

a) The set of pages visited by a specific user at a specific time is called page viewing time. It varies from while default time is 30 minutes by R.Cooley [1].

b) The second method depends on page stay time which is calculated with the difference between two timestamps. If it exceeds 10 minutes the second entry is assumed as a new session[2].

•User Navigation Oriented Heuristics:

a) The set of user sessions are extracted as[1]

USS= {USID, (URI1, ReferURI1, Date1)…..(URIk, ReferURIk,Datek))}

where $1 \leq k \leq n$ and n denotes the amount of records in log file.

b) Another method, define a user session[4] as:$\mathbf{s^{(i)} = (u^{(i)} ,t^{(i)} ,r^{(i)})}$ Where:$u^{(i)} \in U$: is the user identification, $t^{(i)}$: is the access time of the whole session, $r^{(i)}$: is the set of all resources requested during the i-thsession. Summarizing, after the data structuration phase, a set of nsessions $s(i)$ is identified from the log data. We denote the setof all identified sessions by: $\mathbf{S = (s^{(1)} , s^{(2)} ,..., s^{(n_s)})}$.

2. *Data filtering module*After the identification of user sessions, we perform a datafiltering step to remove the less requested resources and retainonly the most requested ones[2]. For each resource r_i, weconsider the number of sessions NS_ithat required theresource r_i, and we compute the quantity $NS_i= \max_{i\ldots nR} NS_i$. Then, we define a threshold ε, and we remove all request with$NS_i<\varepsilon$ are removed.

3. *Data summarization module:* The Data Summarization Module generates reports summarizing the information obtained after the preprocessing of raw data. It provides the necessary information to detect some particular aspects related to theuser browsing behavior or to the traffic of the considered site log file.

7. ***Pattern Discovery***: After formatting the data, the user behavioral patterns are extracted from that data for knowledge discovery. In pattern discovery phase, several data mining techniques are applied to obtain hidden patterns reflecting the typical behavior of users. A variety of machine learning methods have been used for pattern discovery in Web usage mining. These methods represent the four approaches that most often appear in the data mining literature: Statistical analysis,association rule,clustering and classification. In the following, some of these techniques are described.

1. *Statistical analysis*:By analyzing the session file, one can perform different kinds of discriptive statistical analyses (frequency, mean, median, etc) on variables like page view, viewing time and length of navigational path.This statistical information is used to produce a periodic report of the website such as information about user's popular pages, average visiting

time of a page, average time of user's browsing through a site, common entry pages and high-traffic days of site.

2. *Association Rules*: In the process of WUM, once sessions have been identified association rules can be used to relate pages that are most often referenced together in a single server session. Such rules are measured using support and confidence utility.

   Support is a measure based on the number of occurrences of user transactions within transaction logs. The typical rule mined from database is formatted as [9]:

   X→Y [Support,Confidence]…… (1)

   It means the presence of item (page) X leads to the presence of item (page) Y, with [Support]% occurrence of [X,Y] in the whole database, and [Confidence]% occurrence of [Y] in set of records where [X] occurred.

3. Classification: Classification is to build automatically a model that can classify a set of pages. It involves the task of mapping a page into one of several predefined classes. In the Web domain, classification techniques allow one developing a profile of users which are belonging to a particular class or category. It requires extraction and selection of features that based on some demographic information available of users, or based on their access patterns.This technique consists of two steps. The first step is based on the collection of training data set and a model is constructed to describe the features of a set of data classes. The data classes in this step are predefined so it is known as supervised learning. In the second step, the constructed model is used to predict the classes of future data. Typical Algorithms include:Decision trees, Rule-based induction, Neural networks, Genetic algorithms, Bayesian networks

4. *Clustering*:Clustering is another mining technique similar to classification however unlike classification there are no predefined classes therefore, this technique is an unsupervised learning process. This technique is used to group together users or data items that have similar characteristics, so that members within the same cluster must be similar to some extent, also they should be dissimilar to those members in other clusters.In the WUM domain.Clustering of users is to cluster users with similar preference, habits and behavioral patterns. The knowledge that is obtained from clustering in WUM is useful for designing adaptive Websites and designing recommender systems.

8. ***Pattern Analysis***: Pattern analysis is the last step within the overall WUMprocess that has two basic goals. the primary goal is toextract the interesting patterns from theoutput of the pattern discovery method by filtering theirrelative patterns. Another aim is analysis to obtain some key information that offers valuable insights about user's navigational behavior.For example we willunderstand the quantity of users that started from a pageand proceeded through some bound pages and eventuallyvisited their goal page. Also, we will get someinformation concerning page quality or some pages that are good entry points.The mostcommon methodology of pattern analysis is combining WUM tools with a knowledge query mechanism such as SQL or visualization tools.

IV. APPLICATION OF WEB USAGE MINING

The collected web log file can be useful for following application.

- Personalization can be achieved by keeping track of previously accessed pages of user.
- Frequently accessed pages can be used for caching.
- Identifying common access behaviors can be used for site modification.
- Web usage mined patterns can be used to gather business intelligence to identify potential prime advertisement locations.
- Web usage mining can be used in Counter Terrorism and Fraud Detection, and detection of unusual accesses to secure data.

- Web usage mining may be used in network traffic analysis to control traffic.

V. CONCLUSION

This paper gives a more comprehensive idea about theanalysis of web usage mining for browsing behavior of a user.Therefore the design of web pages is veryimportant for the system administrator and web designers. These features have great impact on the number of visitors.So the web analyzer has to analyze with the data of serverlog file for detecting pattern. In this paper we tried to give aclear understanding of the data preparation process andpattern discovery process. Web usage patterns and datamining can be the basis for a great deal of future research.

REFERENCES

[1] R.Cooley. Web Usage Mining: Discovery and Application of Interesting patterns from Web Data.

[2] T. W. Yan, M. Jacobsen, H. Garcia-Molina, and D. Umeshwar.From user access patterns to dynamic hypertext linking. Fifth International World Wide Web Conference, May 1996.

[3] Zhu, J., Hong, J and Hughes, J. “Using Markov Chains for Link Prediction in Adaptive Web Sites,” in Lecture Notes in computer science, 2002, 60-73

[4] G. Castellano, A. M. Fanelli, “log data preparation for mining web usagePatterns”, IADIS International Conference Applied Computing 2007.

[5] Ling Zheng, HuiGui 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.

[6] M. Girvan and M. E. Newman, Community structure in social andbiological networks, Proc. Natl. Acad. Sci. USA 99, 7821 (2002).

[7] M. Jalali, N. Mustapha, A. Mamat, Md N. Sulaiman ,OPWUMP, 2008 An architecture for online predicting in WUM-based personalization system , In the proceedings CSICC conference on Advances in Computer Science and Engineering.

[8] Bettina Berendt, Andreas Hotho and GerdStumme, Towards Semantic Web Mining, 2002 In the Proceedings of the First International Semantic Web Conference on The Semantic Web.

[9] MaryamJafari,ExtractingUsers’Navigational Behavior from Web Log Data: a Survey, Journal of Computer Sciences and Applications, 2013