An Overview of Applying Web Log File for Usage Access Pattern Using Weight Values

Nu Yin Kyaw

Abstract**—** In this paper, we focus on how we spawn exploitable web usage access pattern from the raw web log file for personal recommendation. The users' accesses to web sites are stored in web server logs. However, the data stored in the log files do not present an accurate picture of the users' accesses to the web site. Hence, preprocessing of the web log data is an essential and pre-requisite phase before it can be used for knowledge-discovery or mining tasks. In present days, mining from the raw data is important because of the incredible growth of World Wide Web. Without any help on the system, users find it very difficult to extract useful and relevant information from the huge amount of information and the user may spend more time to get the interested information from the website. These problems can be solved by web usage mining which involves preprocessing, pattern discovery and pattern analysis. In many recommender systems, most of the researchers apply and build ontology system for data and concept relationship to generate positive recommendation in web usage mining. In this paper, we want to implement a system without using ontology technology but generate most users interesting usage access pattern from the raw web log file by assigning weights on the activities, interest period and visiting item counts of the user. The main feature of this system is that it generates user interesting pattern dynamically to the web users in their next access. This system is very useful for understanding the behavior of the users and also improving the web site design too. The performance of this system also discussed in this paper.

Keywords**—**Data Cleaning, Log data, Personalization, Sessions, Weight, Data Preprocessing.

I. Introduction

WEB personalization is described as any action that makes the web experience of a user personalized to their taste[1]. Using user's preferences, it serves customized content to them where preferences are obtained by explicit or passive observation of user's overtime as they interact with the system. Forming of web objects (pages, etc.) and subjects (users) matching between and across objects and determination of the set of actions to be recommended for personalization are the elements of web personalization. A number of approaches exist for web personalization. Web usage mining is an effective approach in which mining techniques are applied to large web repositories to discover user access patterns automatically.

Nu Yin Kyaw, Research Student, University of Technology, Myanmar. (email : nuyinkyaw@gmail.com)

Some of the algorithms that are commonly used in web usage mining are association rule generation, sequential pattern generation, and clustering. The input for the web usage mining process is a log file which is present in a web server for each web site and contains information about the accounting of who accessed the web site, what pages are requested and in what order. According to a survey by Netcraft, the growth of web sites is multiplying day by day. August 2011 results shows that there are approximately 463,000,317 web sites available which has been doubled when compared with August 2010 survey which have 213,458,815. The number of web users has increased to 444.8% in 2010 from 2000 as per the statistics data given by Internet World Stats. When a user access a web page an entry is created in web server's log file. So the log entries are also increasing. There are four stages in web log mining.[2]

- Data Collection
- Data Preprocessing
- Pattern Discovery
- Pattern Analysis

Specifically, Web usage mining is the application of data mining techniques to discover usage patterns from click stream and associated data stored in one or more web servers to cater to the needs of web-based applications. However, the data stored in the various data sources do not present an accurate picture of the pages requested or the identification of the user. Hence data preparation techniques are necessary to transform the raw server logs into a suitable data file for web usage mining. The goal of the web usage mining is to personalize the delivery of web content, to improve user navigation, to improve web design, to access the information in fast manner and to improve customer satisfaction. In our paper we concentrate the web usage mining topic; it is one of the intensive research areas as its potential for personalized services and adaptive web sites. Web usage mining results mainly depends upon the proper preparation of the data from the web logs. The web log data can be collected from Client side, Server side (or) Proxy servers; here in this paper we collected web logs from Server side.

In this paper, we propose a system for improving the web usage technique in a website and also for the users to collect their interested information in a better way. Our system will create a most User Interested Page; it will be created by assigning weights and ranking the weight based upon the count of the number of occurrence of each item and the time the user

stay on each page category which was collected from the web logs in a session for all users .The study of the users' access pattern extracted from the web log files may help the web designer to understand the user behavior, find out the interested object of the website and rearrange the structure and design of the web site based upon it.

We also give more attention for important data preprocessing parts such as: identifying clients and collecting the URLs information from the user sessions for the analysis of HTTP requests made by clients. Since a successful analysis is based on accurate information and quality data, preprocessing plays an important role. This paper will be organized as follows: the second section will introduce related work for this study, the third sections present the proposed techniques to generate user interesting usage access pattern, the fourth section presents the result and discussion of our system, and section five presents the conclusion.

II. Related Work

There are lots of approaches dealing with web usage mining for the purpose of discovering the user access pattern to improve web site design. However, data preprocessing in web usage mining has received less attention than its importance warrants. Robert Cooley, Bamshad Mobasher and Jaidep Srivastava presented methods for user identification, session identification, page view identification, path completion, and episode identification [3].They proposed some heuristics to deal with the difficulties during data preprocessing. Bettina Berendt and her colleagues compared time-based and referrer-based heuristics for visit reconstruction [4].Doru Tanasa and Brigitte Trousse proposed advanced data preprocessing. They offered the possibility of jointly analyzing multiple Web server logs. Pei *et al*. [5] have successfully used the log data from Web logs to discover frequent patterns, they proposed an algorithm called Web Access Pattern (WAP) tree for efficient mining of access patterns from pieces of logs.

Dr.K.Iyakutti and P.Arun [6] also proposed a web personalized system in order to understand the behavior of the users and also to improve web site design. In this model, user identification is considered under the client IP address only. Session identification is considered using predefined time based method. In fact, time based method is not appropriate for session identification. They offered the inaccurate performance and results when giving personalized recommendation to users. However, this model has no severe drawbacks.

III. Gathering and Computing Web Log from User Session for Creating User Interesting Usage Access Pattern

We collect the web logs from the testing commercial website, it has 9 objects and we collect around 30,000 web logs for a period of 60 days. The web log file contains the following entries:

- Client IP address or host name
- Access time
- HTTP request method(GET, POST)
- Path of the resource on the Web server
- Protocol used for transmission
- Status code
- User agent(browser, operating)
- Referrer

From the collected web log data, the following steps have to be done to make the raw web logs to a usable one.

❖ **Data Cleaning:** The raw server log files are unsuitable for access pattern analysis. It is important to remove all the requests from the web log file that are not explicitly requested by the user. When a user requests any page using the browser, there are a number of log entries created in the log file as a page contains other web objects like, images, java script files and cascading style sheets apart from the main HTML page. This can be removed by checking the suffix of the URL name.

In addition to this, erroneous files can be removed by checking the status of the request (such as a status of 404 indicates that the requested file was not found at the expected location). A status with value of 200 represents a succeeded request. A status with value different from 200 represents a failed request.

❖ **User Identification:** Though Web usage analysis does not require the identity of the users, it is essential to differentiate among users. This step requires the identification of unique users. The web usage mining methods that rely on user cooperation are the easiest ways to deal with usage mining problem. However, it may be difficult because of security and privacy. In most cases, the log file provides only the computer address (name or IP) and the user agent. For web sites that require user registration, the log file also contains the user login (as the third record in a log entry) that can be used for the user identification .In our system, the user login is not available for security and privacy; it uses the following heuristics to identify the user. In previous studies, user identification is done under IP address heuristic only. IP addresses, alone, are generally not sufficient for user identification. Our proposed algorithm for user identification will improve the efficiency and the accuracy of the system for personalization. And our system can identify more possible users who visit to our system from one IP address. So, it can generate likely positive recommendations to more users as users may be more than one from one client IP address.

Proposed Algorithm for User Identification

Input: ***N entries of web log file***

Output: ***Identified User Sets***

***Algorithm*:**

```
While (! last entry of log file)
{
Compare IP address of first log entry with IP address of second log entry.
    If (both IP are same)
        Compare the user agent of both entries
        If (both agents are same)
            Check requested page is linked with the previous access pages.
```

If (they are linked)
Identify request entries are from the same user.
Else
Assume that they are different users.
Else // *for different agents*
Assume that they are different users.
Else
Assume that they are different users. /* *IP are different* */
} // while loop

❖ **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 website. A user may have a single or multiple sessions during a period. Once a user has been 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. Sessionization is the process of segmenting the user activity record of each user into sessions, each representing a single visit to the site.

There are three methods in session reconstruction. Two methods depend on time and one on navigation. The simplest methods are time oriented in which one method based on total session time and the other based on single page stay time. The set of pages visited by a specific user at a specific time is called page viewing time. It varies from 25.5 minutes to 24 hours while 30 minutes is the default timeout by Cooley. The second method depends on page stay time which is calculated with the difference between two timestamps. If it exceeds 10 minutes then the second entry is assumed as a new session. Time based methods are not reliable because users may involve in some other activities after opening the web page and factors such as busy communication line, loading time of components in web page, content size of web pages. Third method based on navigation uses web topology in graph format.

The following is the proposed algorithm that will be used to identify user's session in our system .

roposed Algorithm for Session Identification

Input: ***N requests of user set, Visiting Maximum Time, Visiting Minimum Time, 2D array*.**

Output: ***Identified Session Sets***

Algorithm:

While (! Last entry row of 2D array)
{
Step 1: Calculate the visiting time of a web page of a user.
Step 2: Compare the visiting time with Visiting Maximum Time and Visiting Minimum Time of each web page.
If the visiting time is <u>less</u> than Visiting Maximum Time then assign the weight as <u>0</u>.

Else if visiting time is <u>between</u> Visiting Maximum Time and Visiting Minimum Time then assign the weight as <u>1.</u>

Else if visiting time is <u>greater</u> than Visiting Maximum Time then assign the weight as <u>100</u>. And if <u>referrer URL</u> is <u>null</u> then weight is assigned as <u>000</u>.

If the same page is visited by the user again in each user's set then increment the corresponding entry object.

}// while loop

Here, visiting time for a particular page is determined by finding the differences between the time fields of two consecutive entries of a same user. Website Administrators fix the minimum visiting time and maximum visiting time for all web pages as per the contents, loading time of a web page and leaving time of a user for a web page. For example home page will take less time to browse. Our method for session identification can leave out user's uninteresting web pages and impossible web pages for manipulating of web personalization of the users. As more visiting time on a web page may not be user's more interesting web page. So, this proposed algorithm will help more in finding user's most interesting web page as it can leave poor web pages from the possible calculated visiting time on a web page.

Our proposed system define visiting minimum time and maximum time according to **leaving time** of a user for a web page, **loading time** of a web page and **average time probability** for word and image contents of a web page.

Minimum Visiting Time*= (*Minimum Loading Time of a web page*) + (*Minimum Leaving Time of a user for that Web Page*) + (*Minimum Average Time for content on a page)

Maximum Visiting Time*= (*Maximum Loading Time of a web page*) + (*Maximum Leaving Time of a user for that Web Page*) + (*Maximum Average Time for content on a page)

We compare the visiting time of a user with Minimum Visiting Time and Maximum Visiting Time. If visiting time is between Minimum Visiting Time and Maximum Visiting Time, we consider it as usual page. If not, we consider it as irregular page for further processing.

❖ **Web Log categorizing:** It is categorizing of the web logs based upon the items which were accessed by each user. For that, it first categories every item of the website and coded as numeric based sequence.

Code	*HTML Page Category*
1	air-con
2	jewelry
3	artist
4	book

Gathering Click Stream Data: It is collecting of the click stream data for each user from each user session.

Users	*Request Access Pattern*
192.168.1.2	123413231342344341
95.102.3.4	1242431234123

❖ **Counting occurrence / Assigning weight & Ranking order:** Count the number of occurrence of each item. Based upon the count, the activities that the user makes and the time user stay on each page category, it assigns weight and ranking the weblog data in the order of weights**.**

Weight Value for each category$_i$= (*number of page count in each category$_i$ * threshold value*) + (*time spent in each category$_i$ * threshold value*) +
(Activities Weight Values (
$(\sum_{i=1}^{n} ci\ w1 =$ *threshold value for activity one*$)$+
$(\sum_{i=1}^{n} ci\ w2 =$ *threshold value for activity two) +*
$(\sum_{i=1}^{n} ci\ w3 =$ *threshold value for activity three*$)$
) in each category$_i$).

There are ***three activities*** in our system, ***n*** is the number of categories in our system, c_i are categories and ***w1, w2 and w3*** are weight values.

❖ **Generating user interesting access pattern:** Find the interesting access pattern for every user. It generates most user interesting page for personalization based upon their previous access information (web logs).

IV. Result And Discussion Of Our System

There are lot of approaches dealing with web usage mining for the purpose of finding the interesting information (or) automatically discover the user pattern, but in our model the whole process will be divided into three sub process they are: data training which contains the collection of web logs and converts the raw web log data into usable one. In that, counting the items for each user, assign weight and ranking the web log based upon the weight (user interest), generating user interesting usage access page from the respective weight value and provide the positive personalization or recommendation to the user.

Our system was implemented in Java programming language. Here we collect the web logs for 60 days from the testing commercial web site. Our experiments were performed on a 2.8GHz Pentium IV CPU, 512MB of main memory, Windows 2007 professional, MySQL database Server, IIS Server and JDK 1.6.0_02.

Data Cleaning: Data cleaning involves the removal of records with graphics and videos format such as gif, JPEG, etc., and records with robots traversal are removed. In Fig.3, Bar chart 1 represents the initial requests in raw web log. From Bar chart 2 to 6 represent the requests after removing the log entries with filename suffix “gif” or “GIF”, the log entries with filename suffix “jpg” or “jpeg”, the log entries with filename suffix “css”, robots’ requests and error’s requests. The number of records resulted after cleaning phase is 1476 and it is represented in Fig.1.

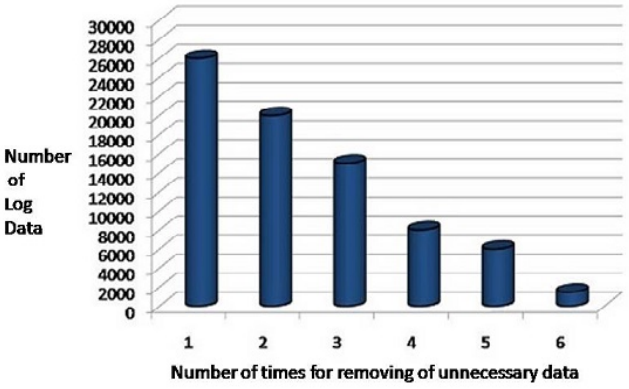

Fig.1. Data Cleaning Process

A. User Identification

After the data cleaning process is performed, users are identified by using IP address and User Agent and Referrer fields. Fig.2 is the processes of user identification. Bar chart 1 is the number of users identified only by only IP addresses. Bar chart 2 is the number of users with the same IP address and agent. Bar chart 3 is the number of users considering under same IP address, agent and referrer.

Fig.2. User Identification Process

B. Session Identification

In most session identification process, it is done by time-based method such as single page stay time and total session time. Time based methods are not reliable because users may involve in some other activities after opening the web page and factors such as busy communication line, loading time of components in web page, content size of web pages are not considered. In this case, irregular web page that has longer usage time for abnormal case and this page is also considered in session part as most important page for positive recommendation decision because of having longer duration time on it. But in our system, such page that are more than possibility maximum browsing time and less than minimum browsing time are not considered in our session. So, we can eliminate such poor web page for further process of generating user interesting usage access pattern. According to the comparison of our system and normal system that use single

page stays time session identification method in the times of 120,180,240 and 300 seconds, our system can generate accurate and correct user interesting web page in any time periods. But, the system that uses time based session identification method can't give accurate result in the time of 180 seconds and more because they consider such web page that have impossible longer duration time period as a more weighted page.

V. Conclusion

Today, web usage mining has emerged as the essential tool for realizing more personalized user-friendly and business optimal web services. Advances in data pre-processing, modeling, and mining techniques, applied to the web data, have already resulted in many successful applications in adaptive information systems, personalization services, web analytics tools, and content management systems. So the usage of data mining methods and knowledge discovery on the web is now on the attention of an enhancing number of researchers. This paper has presented a way of generating user interesting access pattern with promising algorithms for user and session identifications. The discovered patterns can then be used for various web usage applications such as site improvement, business intelligence and effective recommendations for the user. The main aim of our system is to improve the performance of the access method (i.e.) the personalization process will surely improve the system performance compare to the normal access by the user. It can find most user interesting web pages correctly with simple from only web raw log data on a server without building specific ontology system to find relationship and association between pages. The limitation of our system is, to create user interesting usage access page for the new users in the website, because we will create such page from the web log information only. For the new users there is no web log data. But, the results produced by our research can provide guidelines for improving the design of web applications and give positive recommendation to user with greater accuracy than normal case.

References

[1] Areej Al-Qwaqenah, Belal Abu Ata and Mohammed Al-Kabi," Discovering the Web Usage in three Jordanian Universities".

[2] Dimitrios Pierrakos**,** Georgios Paliouras,Christos Papatheodorou, Constantine D. Spyropoulos," KOINOTITES: A Web Usage Mining Tool for Personalization".

[3] Li Chaofeng, "Research and Development of Data Preprocessing in Web Usage Mining".

[4] C.P. SUMATHI, R. PADMAJA VALLI, T. SANTHANAM, "An overview of preprocessing of web log files for web usage mining", *Journal of Theoretical and Applied Information Technology, 15th December 2011*.

[5] J. Pei, J. Han, B. Mortazavi-Asl, H. Zhu, "Mining access patterns the efficiently from web logs" *in PADKK '00: Proceedings of the 4 Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications. London, UK: Springer- Verlag, pp. 396-407, 2000*.
http://dx.doi.org/10.1007/3-540-45571-X_47

[6] P.Arun, K.Iyakutti, "Ontology Generation from Session Data for Web Personalization", *Int. J. of Advanced Networking and Application 241 Volume: 01, Issue: 04, Pages: 241-245 (2010).*

[7] Castellano, A. M. Fanelli, M. A. Torsello, "Log Data Preparation for Mining Web Usage Patterns".

[8] L.K. Joshila Grace, V.Maheswari, Dhinaharan Nagamalai," Analysis of Web Logs and Web User In Web Mining", *International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1, January 2011.*

[9] Chao Liu and colleagues, "Leaving Web pages: The Weibull Hazard Function", *Jakob Nielsen's Alertbox: September 12, 2012 at Microsoft Research.*

[10] Harald Weinreich Hartmut Obendorf, Eleco Herder, and Matthias mayer,"Not Quite the Average:An Empirical Study of Web use", *in the ACM Transactions on the Web,vol.2,no.1 (February 2008), article #5*.

[11] Claudia Elena DINUC," Association and Sequence Mining in Web Usage", *Annals of "Dunarea de Jos" University of Galati Fascicle I. Economics and Applied Informatics Years XVII – no2/2011 ISSN 1584-0409 www.ann.ugal.ro/eco*.

[12] Mr. Sanjay Bapu Thakare, Prof. Sangrarn. Z. Gawali," A Effective and Complete Preprocessing for Web Usage Mining," *(l1CSE) International Journal on Computer Science and Engineering Vol. 02, No. 03, 2010,848-851.*

Nu Yin Kyaw is a Research Student for Information Technology at University of Technology (Yatanarpon Cyber City, Myanmar). Her research interests are Web mining and Software Engineering. Contact her at nuyinkyaw@gmail.com.