DISCOVERING USER IDENTIFICATION MINING TECHNIQUE FOR PREPROCESSED WEB LOG DATA

[1] ASHWIN G. RAIYANI, [2] PROF. SHEETAL S. PANDYA

[1, 2] Department Of Computer Engineering,
[1,2] RK. University, School of Engineering.

Ashwin.rkcet@gmail.com, sheetal.rkcet@gmail.com

***ABSTRACT*:** *The Web Usage Mining can be described as the discovery and Analysis of user access pattern through mining of log files and associated data from a particular websites. No. of visitors interact daily with web sites around the world. Enormous amount of data are being generated and these information could be very prize to the company in the field of accepting Customer's behaviors. In this paper a complete preprocessing style having data cleaning, user and session Identification activities to improve the quality of data. Enhanced preprocessing technique one of the User Identification which is key issue in preprocessing technique phase is to identify the unique web users. Traditional User Identification is based on the site structure, being supported by using some heuristic rules, for use of this reduced the efficiency of user identification solve this difficulty we introduced proposed Technique DUI (Distinct User Identification) based on IP address ,Agent and Session time ,Referred pages on desired session time. Which can be used in counter terrorism, fraud detection and detection of unusual access of secure data, as well as through detection of regular access behavior of users improve the overall designing and performance of upcoming access of preprocessing results.*

KEYWORDS— Preprocessing, Server log, Session time, User Identification, Web Usage mining.

I: INTRODUCTION

Web Mining is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World Wide Web. Web usage mining provides the support for the web site design, providing personalization server and other business making decision, etc. In order to better serve for the users, web mining applies the data mining, the artificial intelligence and the chart technology and so on to the web data and traces users' visiting characteristics, and then extracts the users' using pattern[l]. It has quickly become one of the most important areas in Computer and Information Sciences because of its direct applications in e-commerce, CRM, Web analytics, information retrieval and filtering, and Web information systems. According to the differences of the mining objects, there are roughly three knowledge discovery domains that pertain to web mining: Web Content Mining, Web Structure Mining, and Web Usage Mining. Web content mining is the process of extracting knowledge from the content of documents or their descriptions. Web document text mining, resource discovery based on concepts indexing or agent; based technology may also fall in this category. Web structure mining is the process of inferring knowledge from the World Wide Web organization and links between references and referents in the Web.

Web UsageMining

Finally, web usage mining, also known as Web Log Mining, is the process of extracting interesting patterns in access logs. Web Usage Mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications [4]. The results of Web Usage Mining can be used in personalization, system improvement, site modification, business intelligence, usage characterization and so forth.

Generally, Web Usage Mining consists of three processes: Data Preprocessing, Patterns discovery and Patterns analysis. As the data sources of patterns discovery, the results' quality of data preprocessing influences the results of patterns discovery directly. Better data sources can not only discover high quality patterns but also improve the algorithm of Web Usage Mining. So, data preprocessing is particularly important for the whole Web Usage Mining processes and the key of the Web Usage Mining's quality. However, Each type of data collection used in data preprocessing differs not only in the terms of the location of the data source, but also the kinds of data available, the segment of population from which the data are collected, and it's method of implementation. The research on data preprocessing of Web Usage Mining is a focus field nowadays. These attempts to present the process of data preprocess in data preprocessing of Web Usage Mining.

II:PREPROCESSING TECHNIQUE

Ideally, the input for the Web Usage Mining process is a user session file that gives an exact account of who accessed the Web site, what pages were requested and in what order, and how long each page was viewed. A user session is the set of the page accesses that occur during a single visit to a Web site. However, because of the reasons we will discuss in the following, the information contained in a raw Web server log does not reliably represent a user session file before data preprocessing. Generally, data preprocessing consists of data cleaning, user identification, session identification and path completion, as shown in Figure -2.

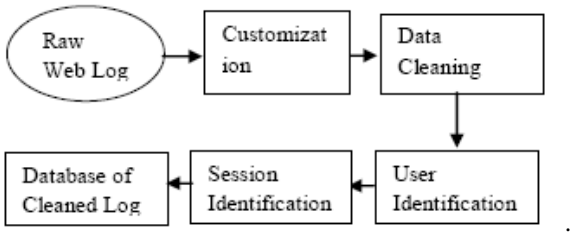


Fig 2: Preprocessing Technique

1) Data Cleaning:

The purpose of data cleaning is to eliminate irrelevant items, and these kinds of techniques are of importance for any type of web log analysis not only data mining.According to the purposes of different mining applications [4], irrelevant records in web access log will be eliminated during data cleaning. Since the target of Web Usage Mining is to get the user's travel patterns, following two kinds of records are unnecessary andshould be removed:

- The records of graphics, videos and the format information The records have filename suffixes of GIF, JPEG, CSS, and so on, which can found in the URI field of the every record;
- The records with the failed HTTP status code. By examining the Status field of every record in the web access log, the records with status codes over 299 or under 200 are removed

2) User Identification:

The task of user and session identification is find out the different user sessions from the original web access log. User's identification is, to identify who access web site and which pages are accessed. The goal of session identification is to divide the page accesses of each user at a time into individual sessions. A session is a series of web pages user browse in a single access. The difficulties to accomplish this step are introduced by using proxy servers, e.g. different users may have same IP address in the log. A referrer-based method is proposed to solve these problems in this study. The rules adopted todistinguish user sessions can be described as follows:

- The different IP addresses distinguish different users;
- If the IP addresses are same, the different browsers and operation systems indicate different users; User identification. In this step the unique users are distinguished, and as a result, the different users are identified. This can be done in various ways like using IP addresses, cookies, direct authentication and so on. Because the focus of this paper is put on the analysis of the different user identification methods, this step will be discussed later in detail.
- Consider, for instance, the example of Fig. 3. On the left, the figure de-picts a portion of a partly preprocessed log file (the time stamps are given as hours and minutes only). Using a combination of IP and Agent fields in the log file, we are able to partition the log into activity records for three separate users (depicted on the right).

Time	IP	URL	Ref	Agent
0:01	1.2.3.4	A	-	IE5;Win2k
0:09	1.2.3.4	B	A	IE5;Win2k
0:10	2.3.4.5	C	-	IE6;WinXP;SP1
0:12	2.3.4.5	B	C	IE6;WinXP;SP1
0:15	2.3.4.5	E	C	IE6;WinXP;SP1
0:19	1.2.3.4	C	A	IE5;Win2k
0:22	2.3.4.5	D	B	IE6;WinXP;SP1
0:22	1.2.3.4	A	-	IE6;WinXP;SP2
0:25	1.2.3.4	E	C	IE5;Win2k
0:25	1.2.3.4	C	A	IE6;WinXP;SP2
0:33	1.2.3.4	B	C	IE6;WinXP;SP2
0:58	1.2.3.4	D	B	IE6;WinXP;SP2
1:10	1.2.3.4	E	D	IE6;WinXP;SP2
1:15	1.2.3.4	A	-	IE5;Win2k
1:16	1.2.3.4	C	A	IE5;Win2k
1:17	1.2.3.4	F	C	IE6;WinXP;SP2
1:26	1.2.3.4	F	C	IE5;Win2k
1:30	1.2.3.4	B	A	IE5;Win2k
1:36	1.2.3.4	D	B	IE5;Win2k

User 1

0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

User 2

0:10	2.3.4.5	C	-
0:12	2.3.4.5	B	C
0:15	2.3.4.5	E	C
0:22	2.3.4.5	D	B

User 3

0:22	1.2.3.4	A	-
0:25	1.2.3.4	C	A
0:33	1.2.3.4	B	C
0:58	1.2.3.4	D	B
1:10	1.2.3.4	E	D
1:17	1.2.3.4	F	C

Fig. 3: Example of User Identification using IP+Agent

3) 3) Session identification.

A session is understood as a sequence of activities performed by a user when he is navigating through a given site. To identify the sessions from the raw data is a complex step, because the server logs do not always contain all the information needed.

There are Web server logs that do not contain enough information to reconstruct the user sessions, in this case (for example time-oriented or structure-oriented) heuristics can be used as describe

- If all of the IP address, browsers and operating systems are same, the referrer information should be taken into account. The Refer URI field is checked, and a new user session is identified if the URL in the Refer URI field hasn't been accessed previously, or there is a large interval (usually more than 10 seconds) between the accessing time of this record and the previous one if the Refer URI field is empty;

- The session identified by rule 3 may contains more than one visit by the same user at different time, the time oriented heuristics is then used to divide the different visits into different user sessions. After grouping the records in web logs into user sessions, the path completion algorithm should be used for acquiring the complete user access path.

The WUM system presented in this paper is not a full web log mining system. Its aim is to better identify web users and individuals behind the users. In this manner it realizes the first three steps of a web log mining process. The results provided by our system can be used for further processing by any data mining algorithm.

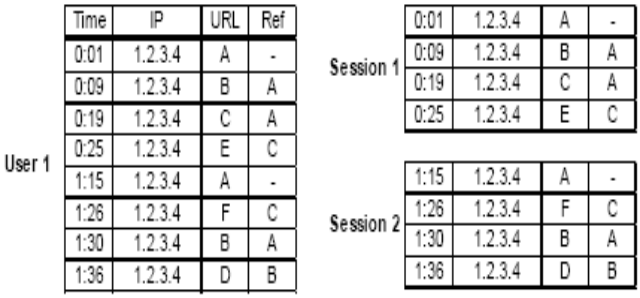
A sessionization heuristic h at-tempts to map R into a set of constructed sessions, denoted by Ch. For the ideal heuristic, h*, we have Ch* = R. In other words, the ideal heuristic can re-construct the exact sequence of user navigation during a session. Generally, sessionization heuristics fall into two basic categories: time-oriented or structure-oriented. Time-oriented heuristics apply either global or local time-out estimates to distinguish between consecutive sessions, while structure-oriented heuristics use either the static site structure or the im-plicit linkage structure captured in the referrer fields of the server logs. Various heuristics for sessionization have been identified and studied [5]. More recently, a formal framework for measuring the effectiveness of such heuristics has been proposed, and the impact of different heuristics on various Web usage mining tasks has been analyzed [7].

As an example, two variations of time-oriented heuristics and a basic navigation-oriented heuristic are given below. Each heuristic h scans the user activity logs to which the Web server log is partitioned after user identification, and outputs a set of constructed sessions:

• ***hur1***: Total session duration may not exceed a threshold θ. Given t_0, the timestamp for the first request in a constructed session S, the request with a timestamp t is assigned to S, iff $t - t_0 \leq \theta$.

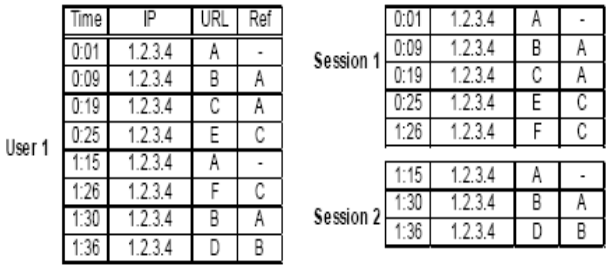
• ***hur2*:**Total time spent on a page may not exceed a threshold δ. Given t_1, the timestamp for request assigned to constructed session S, the next re-quest with timestamp t_2 is assigned to S, iff $t_2 - t_1 \leq \delta$.

• ***hur*-ref**: A request q is added to constructed session S if the referrer for q was previously invoked in S. Otherwise, q is used as the start of a new constructed session. Note that with this heuristic it is possible that a re-quest q may potentially belong to more than one "open" constructed session, since q may have been accessed previously in multiple sessions. In this case, additional information can be used for disambiguation. For example, q could be added to the most recently opened session satisfying the above condition



Time	IP	URL	Ref
0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Fig. 4 Example of sessionization with a time-oriented heuristic



Time	IP	URL	Ref
0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Fig. 5 Example of sessionization with the h-ref heuristic

An example of the application of sessionization heuristics is given in Fig. 5 and Fig. 6. In Fig. 5, the heuristic h_1, described above, with θ = 30 minutes has been used to partition a user activity record (from the example of Fig. 4) into two separate sessions.

If we were to apply h_2 with a threshold of 10 minutes, the user record would be seen as three sessions, namely, A->B->C->E, A, and F->B->D. On the other hand, Fig. 5 depicts an example of using h-ref heuristic on the same user activity record. In this case, once the request for F (with time stamp 1:26) is reached, there are two open sessions, namely A->B->C->E and A. But F is added to the first because its referrer, C, was invoked in session 1. The request for B (with time stamp 1:30) may potentially belong to both open sessions, since its referrer, A, is invoked both in session 1 and in session 2. In this case, it is added to the second session, since it is the most recently opened session.

Episode identification can be performed as a final step in pre-processing of the clickstream data in order to focus on the relevant subsets of page-views in each user session. An **episode** is a subset or subsequence of a session comprised of semantically or functionally related page views. This task may require the automatic or semi-automatic classification of page-

III:WEB USER IDENTIFICATION – RELATED WORK

The quantity and quality of the data used in WUM. The Web usage data available for analysis can reach several GB per hour, therefore, septic preprocessing

methods need to be designed to unused data and to structure the raw data.

Our approach provides a complete preprocessing methodology for WUM that will allow the analyst to transform any collection of Web server log into a structured collection of Web requests. Thus, our preprocessing will allow the Inter sites WUM. We divided the preprocessing stage in four main steps: data fusion, data cleaning, Data structuration and data summarization. Cleaned by removing all unnecessary requests, such as implicit requests for the objects embedded in the Web pages. Then, the remaining requests are grouped by user, user sessions, page views, visits and episodes. Ideally, this collections of requests is saved into a relational database with a model that captures the new structure (obtained after the preprocessing) of the Web requests collection.

1. *Problem at time of User Identification:*

User's identification is, to identify who access Web site and which pages are accessed. If users have login of their information, it is easy to identify them. In fact, there are lotsof user do not register their information. What's more, there are great numbers of users access Web sites through, agent, several users **use** the same computer, firewall's existence, one user use different browsers, and **so** forth. **All** of problems make this task greatly complicated and very difficult, to identify every unique user accurately. We may use cookies to track users' behaviors. But considering individual privacy, many users do not use cookies, so it is necessary to find other methods to solve this problem. For users who use the same computer or use the same agent, how to identify them?

As presented in [3], it uses heuristic method to solve the problem, which is to test if a page is requested that is not directly reachable by a hyperlink from any of the, pages visited by the user, the heuristic assumes that there is another user with the same computer or with the same IP address. Ref. [4] presents a method called navigation patterns to identify users automatically. But all of them are not accurate because' they only consider a few aspects that influence the process of users identification. Considering this actuality, we presented a new algorithm called**"**DUI (DISTINCT USER IDENTIFICATION)**"**. It analyses more factors, such as user's IP address, Web site's topology, browser's edition, operating system and referrer page. This algorithm possesses preferable precision and expansibility. It can not only identify users but also identify session. Session identification will be discussed in next section.

The success of the web site cannot be measured only by hits and page views. Usually, the need to know how users behave comes later. Web site design should enclose techniques, which relate web pages and access activity into Standard database format and make data preparation process easier. Unfortunately, web site designers and web log analyzers do not usually cooperate. This causes problems such as identification unique users, construction discrete user's sessions and collection essential web pages for analysis. The result of this is that many web log mining tools have been developed and widely exploited to solve these problems.

However, as will be shown in this research, neither commercial, neither free tools solve adequately these problems. Several steps called ***knowledge discovery***must be passed through in order to observe patterns from data. These steps are (a) data pre-processing which includes such stages as data cleaning, feature selection, transformation, (b) data analysis and (c) finally results visualization and examination.

The role of this preprocessing is to considerably reduce the large quantity of Web usage data available and, at the same time, to increase its quality by structuring it and providing additional aggregated variables for the data mining analysis that will follow. Based on the researches that were already conducted in this domain, we split the data preprocessing step in two main parts: the classical data preprocessing, where we group the methods commonly used in the literature for preprocessing data, and the advanced data preprocessing containing new ideas for data enhancement for the data mining steps that follow.The Classical data preprocessing involves three steps: data fusion, data cleaning, anddata structuration. Our solution for WUM preprocessing also adds what we call anadvanced data preprocessing step (**This consists in a data summarizationstep**, which will allow the analyst to select only the information of interest.

2. Distinct User Identification

About Users identification and Session identification. There are a lot of sequential records come from the same IP address in Web log files [5]. If we use algorithm to check every record mentioned above, it will decrease algorithm's efficiency. If the current record's IP address is the same as previous record's, then we assume thatthe two record come from the same user. Now some definitions are given.

Definition 1: Usersi=(User_ID, User-IP, User-Url,User-Time, User-Referer-Page, User-Agent), O<i<n, where n is the number of total users; User-ID is users' ID have beenIdentified; User-IP is user's IP address; User-Url**is** Web pages user ' accessed; User-Time is time user accessed, User-Referer-Page is the last page the user requested; User-Agent is agent user used.

Input: N no of records of web log file, each record including IP,Agent. IP
Output: Distinct User set identified

a) While(i<N) where (1<=i<=n)
b) {
c) If(DUi.IP != DUi+l.IP) //whether IP is the same.

d) { Theuesr is a new one. }
e) Elseif(DUi.Agent != DUi+ I.Agent)
//Judege the browser and operating system
j) { The user is a new one. }
g) Elseif(DUi.URL has been requested or the referred page of DUi.URL is null)
h) { The user is a new one.}
i) Else
j) { The user is the same one.}
k) i = i+l;
I) }

DISTICT USER IDETIFICATION ALORITHM

Definition: given a clean and filtered web log file and record set web log file R= {r1,r2,r3……r.n} where n>0
Step1: input LogdatabaseRUserof N records
Step2: Distinct User identification base
Step3: RUser=P<url, ip, agent, method, operating system, status,sessionid,time_stamp>
Step4: RUSer=<r1,r2,r3…rn> where n!=0
Step5: for i=1 to n
Step6: read LogdatabaseRUser
Step7 check **if *r(i).userip*** *not part of Distinct user identification base then it treated as new user and copy userip in distinct user identification base.*
Step8: end if
Step9:end loop
Setp10:end

Fig 3. Proposed Technique for user identification (DUI Algorithm)

IV:IMPLEMENTING OUR METHODOLOGY

To support our methodology, we designed and implemented AxisLog Miner. This software tool takes as input the different log files and outputs a MySQL database. For more details on our WUM relational model and the preprocessing tool[7], We use Perl scripts to implement data fusion, data cleaning, and data structuration. For the user interface and data summarization, we use Java and SQL.The site map module is under development and will be soon integrated in the software tool. More advanced are On-line Analytical Processing (OLAP) tools which give analyst a multidimensional view of the data. An important role is played also by visualization tools presenting the interesting results in graphs and charts.

Entries in raw web log	**47890**
Entries after data cleaning	**12783**
Number of users	**6542**
Number of Unique users	**4366**
Number of sessions	**6744**

Transfer Server Logs to database

After reading the log files, several attributes are ignored because they were considered not important for the analysis. The read logs records will be stored in a database. Fig. 5shows the database to store the data.

1 Declare Variables
2
3*SetDBconn=Server.CreateObject("ADODB.Connection ")*
4*SetRSconn= Server.CreateObject("ADODB.Recordset")*
5
6 *ConnStrng = {MsAccess Driver}*
7 *DB.OpenConnStrng*
8 *RS.OpenTableName, ActiveConnection,*
9
10*Add Datafiled*
11*RS.Update… …*
12*Set Rsconn = Nothing ……*
13*DBconn.Close*

Fig. 5 : Algorithm transfer to database

V: CONCLUSION

In this Research we present Distinct user identification technique which enhancement of pre-processing steps of web log usage data in data mining. We use two pre-processing technique combine within one pre-processing step time of user identification we find out distinct user based on their attended session time. I introduced one proposed algorithm for advanced pre-processing DUI algorithm is very efficient as compare to other identification techniques. We get more precious accurate result. Based on this we can easily personalized websites, improve the design of WebPages. As usages of users on websites. Future work needs to be done to combine whole process of WUM. A complete methodology covering such as pattern discovery and pattern analysis will be more userful in identification method..

REFERENCES

[1] Ramya C, Dr. Shreedhara K S and Kavitha G, Preprocessing: A Prerequisite for Discovering Patterns in Web Usage Mining Process, 2011 International Conference on Communication and Electronics Information (ICCEI 2011), 978-1-4244-9481-1/11/$26.00 C 2011 IEEE

[2] TheintTheintAye , Web log Cleaning for mining of web usage patterns, 978-1-61284-840-2/11/2011 IEEE.

[3] MohdHelmyAbd,MohdNorzali, Data Preprocessing on Web Server log for Generalized Association Rule Mining. World Academy of Science, Engineering and technology,48 2008

[4] DeMin Dong, Exploring on Web Usage Mining and its Application , 5th world Congress on Intelligent Control and Automation, June 15-19,2004,China

[5] V.Chitraa ,Dr.AntonySelvadossThanamani A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing , International Journal of Computer Applications (0975 – 8887) Volume 34– No.9, November 2011

[6] Mr. Sanjay BapuThakare, Prof. Sangram. Z. Gawali A Effective and Complete Preprocessing for Web Usage Mining , (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 03, 2010

[7] RenátaIváncsy, and SándorJuhász, Analysis of Web User Identification Methods, World Academy of Science, Engineering and Technology 34 2007.