A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing

V.Chitraa
Assistant Professor
CMS College of Science and Commerce
Coimbatore, Tamilnadu, India

Dr.Antony Selvadoss Thanamani
Reader in Computer Science
NGM College Pollachi,
Coimbatore, Tamilnadu, India

ABSTRACT
The growth of World Wide Web is incredible as it can be seen in present days. Users find it very difficult to extract useful and relevant information from the huge amount of information. The problems can be solved by Web Usage Mining which involves preprocessing, pattern discovery and pattern analysis. Preprocessing is an important process which converts raw web log data into transactions. Application of mining techniques to group user's behavior for personalization is effectively done on transactions constructed from sessions. Sessionization is the identification of sessions and is defined as a set of pages visited by the same user within the duration of one particular visit to a web-site. In this research paper, a new technique for identifying sessions is being proposed for extraction of user patterns. The experimental results show that the proposed Session Identification technique is an effective one to construct sessions accurately.

General Terms
Information Retrieval

Keywords
Data Cleaning**,** Log data**, Personalization**, Reference length, Sessions, Weight

.

1. INTRODUCTION
Web Personalization is described as any action that makes the web experience of a user personalized to their taste. 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. Modeling 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 [1] 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. 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

- Data Collection: Logs which are scattered in different web servers are collected[4].
- Preprocessing: Log file consists of lot of irrelevant entries which is to be removed. To enhance the efficiency of mining noise is to be removed before mining. A series of process like data cleaning, user identification, session identification, path completion and transaction identification are handled.
- Pattern Discovery: Application of various data mining techniques to processed data like statistical analysis, association, clustering, pattern matching and so on.
- Pattern Analysis: Uninteresting rules are ruled out and analysis is done using knowledge query mechanism such as SQL or data cubes to perform OLAP operations.

Among the four stages Preprocessing is an important step because of the complex nature of the web architecture and it takes 80% of mining process. The raw data is then pretreated to get reliable sessions for efficient mining. It includes tasks such as data cleaning, user identification, session identification, transactions construction. Data Collection is the first step in web usage mining process. It consists of gathering the relevant web data from sources like Web servers, Cookies from client browsers and Explicit User Input through registration forms and from Proxy servers. Data Cleaning is the process of removing irrelevant items such as jpeg, gif, sound files and references due to spider navigation to improve the quality of analysis. User Identification is the process of identifying users by using Padres and user agent fields of log entries. A user session is considered to be all of the page accesses that occur during a single visit to a Web site. In Session Identification various methods are used to find set of pages visited by a user within the duration of a particular visit. At last transactions are constructed which are defined as a subset of user session having homogenous pages. Specifically, there are a number of difficulties involved in cleaning the raw server logs to eliminate outliers and irrelevant items, reliably identifying unique users and user sessions within a server log, and identifying semantically meaningful transactions within a user session. This paper presents data preparation techniques and algorithms that can be used in order to convert raw Web server logs into user sessions in order to perform Web Mining. The specific contributions include discussion of heuristics that can be used

to identify Web site users, user sessions, and page accesses that are missing from a Web server log.
The proposed technique deals with session identification in which browsing time of individual users and traversals between pages are considered. Sessions thus formed are applied with mining techniques in the pattern discovery. The remainder of this paper is organized as follows: Section 2 presents various works which had been done previously. Section 3 deals with the proposed technique. Experimental results are given in Section 4. At last a summary of the work is given in Section 5.

2. RELATED WORK

The focus of literature review is to study, compare and contrast the available preprocessing techniques. Data Cleaning is done to remove the invalid records with unsuccessful status, auxiliary entries with image files, robot navigation entries [15]. Users are identified to analyze user behavior. In the log files authenticated information is available for registered websites. But in most of the cases these fields are empty due to user's reluctance to use those sites. Cookies from client side are used for identification. But it is not always possible since users might disable cookies for privacy concern. So the fields used to identify users are IPaddress, user agent and site topology is also checked to identify a new user by the use of links. If the requested page is not reachable from any of the pages visited by the user then the user is identified as a new user in the same address[11].

2.1 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. 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 [3] to 24 hours [13] while 30 minutes is the default timeout by Cooley [11]. 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 are not considered. Third method based on navigation uses web topology in graph format. It considers webpages connectivity, however it is not necessary to have a hyperlink between two consecutive page requests. If a web page is not connected with previously visited page in a session, then it is considered as a different session. Cooley proposed a referrer based heuristics on the basis of navigation in which referrer URL of a page should exists in the same session. If no referrer is found then it is a first page of a new session.

Both the methods are used by many applications. To improve the performance different methods were devised on the basis of Time and Navigation Oriented heuristics by different researchers. Different works were done by researchers for effective reconstruction of sessions. The referrer-based method and time-oriented heuristics method are combined to accomplish user session identification in [4]. A simple algorithm is devised by Baoyao Zhou [2]. An access session is created as a pair of URL and the requested time in a sequence of requests with a timestamp. This algorithm is suitable when there are more number of URL's in a session. The default time set by author is 30 minutes per session. Smart Miner is a new method devised by Murat Ali and team [8]. This framework is a part of their Web Analytics Software. The sessions constructed by SMART-SRA contains sequential pages accessed from server-side works in two stages and follows Timestamp Ordering Rule and Topology rule[9]. Another effective method using Integer Programming was proposed by Robert F.Dell [12] in which all sessions is constructed simultaneously. To identify users and sessions a Cube is also used effectively by storing User ID, IP address, user's URL and time visited in an intelligent algorithm [16].

Graphs are also used for session identification. It gives more accurate results for session identification. Web pages are represented as vertices and hyperlinks are represented as edges in a graph. Traversal is a sequence of consecutive web pages on a base graph traversed by the user. A method was proposed by Mehdi Heydari and team in which client side data is also considered to reconstruct user's session [7]. Graph mining methods constructs accurate sessions and the time taken is also comparatively less. Another algorithm has been proposed by Junjie Chen and Wei Liu in which data cleaning and session identification is combined [14].

2.2 Path Completion

Due to local cache, agent cache ,"POST" technique and usage of browser's 'BACK' button in web pages some important accesses may not recorded in the log file. To find the missing pages path completion step is necessary. Path Completion is done by analyzing URLs and referrer URLs in a user's session. If a page request made is not directly linked to the last page requested, the recent history of session is searched and if the page is available previously as referrer URL, the related URL of the previous entry is added in path. Completed paths give a complete navigation of a user in a particular visit.

2.3 Construction of Transactions

The goal of transaction identification is to create meaningful clusters of references for each user [11]. Transaction identification is done by merges or divides approaches. To find out the user's travel pattern and user's interests, two kinds of transactions are present, travel path transactions and content only transactions. Travel path transaction is a combination of auxiliary and content pages accessed by a user. The content only transactions are only content pages which are used in mining to discover user's interest and cluster users visiting the same web site[10].

3. PROPOSED METHOD

3.1. Web Usage Data

Whenever a user hits a page the log data is collected automatically in Web servers. It represents the accurate navigational behavior of visitors. It is the primary source of data in Web usage mining. Each hit against the server, corresponding to an HTTP request, generates a single entry in the server access logs. There are different forms of log files like Apache, IIS etc., Each log entry may contain fields such

as date time 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 sc-bytes cs-bytes. A sample log is given below

```
2007-12-06 05:22:16 ::1 GET /iisstart.htm - 80 - ::1
Mozilla/4.0+(compatible;+MSIE+7.0;+Windows+NT+6.0;+SLCC1;+
.NET+CLR+2.0.50727;+Media+Center+PC+5.0;+InfoPath.1;+.NET+
CLR+1.1.4322;+.NET+CLR+3.5.21022;+.NET+CLR+3.0.04506)
200 0 0 296 336
```

3.2 Data Cleaning

The task of data cleaning is to remove the irrelevant and redundant log entries for the mining process. There are three kinds of irrelevant or redundant data to be removed [5]. They are.

a. Additional Requests: A user's request to view a particular page often results in several log entries. Graphics and scripts are downloaded in addition to the HTML file, because of the connectionless nature of the HTTP protocol. Since the main intention of Web Usage Mining is to get a picture of the user's behavior, it does not make sense to include file requests that the user did not explicitly request. Suffix part of an URL is checked and eliminates suffixes like gif, jpg, GIF, JPEG, css, map etc.

b. Robots' requests: Web robots are software tools that scan a Web site to extract its content. Spiders automatically follow all the hyperlinks from a Web page. To remove robots' request, we can look for all hosts that have requested the page "robots.txt", which is checked by robot while browsing.

c. Entries with error: Status code shows the success or failure of a request. Entries with status code less then 200 and greater than 299 are failure entries which are to be removed.

Only necessary fields like date, time, IPaddress, User Agent, URL requested, URL referred, time taken are considered for further experiments to reduce the processing time So attribute subset selection is done.

3.3 User Identification

The strategy of User Identification based on the log entries without considering the topology structure of site. The description of concrete strategy algorithm is as follows.

Input: N records of web log file
Output: User sets identified
Algorithm:
Repeat steps
1: Compare ipaddress of first log entry with ipaddress of second log entry.
2: If both are same compare the user agent of both entries else assume as different users.
3:If both user agents are same identify both entries are fromsame user.
until last entry
User's IP addresses of two consecutive entries are compared. If the IP address is the same, user's browser and operating system is verified and if both are same, both the records are considered from the same user. These experiments prove that the algorithm significantly improves the efficiency and the accuracy of user identification without usage of site topology.

3.4 New Session Identification

In the proposed method a matrix is constructed from which sessions are identified. User's traversal is the input to the method. Matrix consists of rows and columns in which columns are the web pages and rows are users and their sessions identified from the above algorithm. Browsing time for a particular page BT is determined by finding the differences between the time fields of two consecutive entries of a same user. In IIS 7.0 time-taken is another field which is the processing time of the server. So this time is also deducted from the Browsing time.

3.4.1 Approximate Browsing time

Website Administrators fix the minimum time and maximum time for all web pages as per the contents. For example home page will take less time to browse. This arbitrary time fixed by web site designer is designated as BTmin and BTmax [6]. Compare the browsing time BT of a user with BTmin and BTmax. Weights are calculated which are numeric values to store in the matrix cells. The advantage of codification of weights is that the behavior of users such as navigation pages , interested pages, longer duration pages is also known accurately for grouping them while discovering patterns. The codes for weights are fixed from 0,1 to 9, 10,100.

If BT <BTminwt = 0
Else If BT >BTmin and BT <BTmax then wt = 1 to 9
Else if BT >BTmax then wt =10
Else if referURL == null then wt = 100.

1 to 9 is the valid Browsing time, 10 is the longer time taken by user and if referrerURL field is null a weight 100 is assumed. Valid Browsing Time is considered as a range because of the Cache memory problem. Whenever a user uses 'BACK' key to view visited page, and if the page is not having an 'expires' property within a short duration, a copy of webpage which is available in client side is displayed and so it's entry does not reach the log file. So it is expected that the same page may be traversed and an entry is stored in log file in 'BACK'. In this regard the weight is incremented in Matrix. So a range is fixed and with an assumption that not more than nine times the same page is revisited in same session.

3.5 Session Construction

The proposed algorithm for session construction is as follows :
Input: User sets with N records, BTmin, BTmax, 2D matrix
Output: Constructed Sessions
Algorithm:
Repeat steps
1: Calculate the browsing time of a web page by a user by finding the difference between two consecutive entries and subtract the time taken value
2: Compare the browsing time with minimum and maximum time of each web page
3: If the browsing time is less than minimum time fix the weight as '0' else if it is between minimum and maximum, then weight is fixed as '1' , if the weight exceeds maximum fix as 10 and if referrer URL is null weight is fixed as 100.
4: If the same page is visited by the user again in user's set increment the corresponding entry.
5: Weights are stored in the matrix in the corresponding cells. The value a_{ij} represents a weight based on users browsing time in $page_j$. until last row in users set.

Each row indicates user's sessions. After the users traversals are found and browsing time is calculated, weights are found. The weights are stored in the matrix cells. It the values stored is 100 the next entry will be stored as next session in next row. Each row is considered as a user's individual session traversal.

4. EXPERIMENTAL RESULTS

The proposed web log preprocessing is evaluated in this section. The results of different stages are given as follows. The method is implemented by using MATLAB.

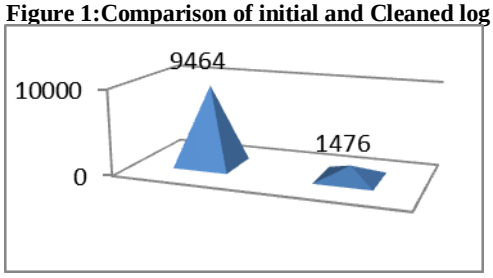
4.1 Web log data

The weblog data considered for evaluation is collected from reputed charitable trust web server during the period of May to August, 2011. Initially the log file consists of 9464 raw log entries with noisy entries like gif, jpeg etc which are not necessary for web log mining.

4.2 Data Cleaning

So data cleaning is performed to remove the unnecessary log which will reduce the processing in determining the web usage pattern. This cleaning phase involves the removal of records with graphics and videos format such as gif, JPEG, etc., and records with robots traversal is also removed. The number of records resulted after cleaning phase is 1476 and it is represented in figure 1.

Figure 1:Comparison of initial and Cleaned log



4.3 Users Identification

After the data cleaning process is performed, users are identified by using IPaddress and UserAgent fields. There are 124 unique users identified after applying the algorithm.

4.4 Sessions Identification

Session identification process is carried out. A part of the result obtained by using the session identification process is represented in matrix format is presented in figure 2. The values in the matrix represent the appropriate weight which is determined by the proposed approach. User 1, User2, etc., represents the user with particular IP address. S1,S2 etc., gives the session constructed as per proposed method. As presented in the figure, it can be seen that for user 1 there are two sessions which has the URL traversal 1-2 and 1-7. For user 2, only one session is resulted i.e., 5-6-7-8. For user 3, the resulted sessions are two with URL traversals such as 2-4 and 7-8-9. For user 4, the resulted session is 1 with a traversal 4-5-10. The traversal of URL 2-6 is resulted for the user 5. For the user 6, the resulted sessions are 4-6 and 9-10. For user 7, the session identified has 1-4-7 traversal. These sessions will helps in determining the significant URL in the web site.

Figure 2: Sample Matrix for Session Identification

Users	Sessions	URL's									
		1	2	3	4	5	6	7	8	9	10
U1	S1	1	2	100							
U1	S2	2	0	0	0	0	0	2	100		
U2	S1	0	0	0	0	2	1	1	2		
U3	S1	0	1	0	2	100					
U3	S2	0	0	0	0	0	0	1	1	3	
U4	S1	0	0	0	2	1	0	0	0	0	1
U5	S1	0	4	0	0	0	1				
U6	S1	0	0	0	3	0	5				
U6	S2	0	0	0	0	0	0	0	0	1	10
U7	S1	2	0	0	1	2	0	1	100		

5. CONCLUSION

The growth of the web has outcome in a huge amount of information that is now freely offered for user access. The several kinds of data have to be handled and organized in a manner that they can be accessed by several users effectively and efficiently. So the usage of data mining methods and knowledge discovery on the web is now on the spotlight of a boosting number of researchers. Web usage mining is a kind of data mining method that can be useful in recommending the web usage patterns with the help of users' session and behavior. Web usage mining includes three process, namely, preprocessing, pattern discovery and pattern analysis. This paper mainly focused on preprocessing approach. In the data cleaning phase, unnecessary records including graphics files, robots are removed. The next process is the identification of sessions which is derived by forming the user behavior in matrix format. Experimental results suggest the significance of the proposed approach.

6. REFERENCES

[1] Bamshad Mobasher, Robert Cooley, Jaideep Srivastava, "Automatic Personalization Based on Web Usage Mining**",** Communications of the ACM, New York, Volume 43, Issue 8, Aug 2000.

[2] Baoyao Zhou, Siu Cheung Hui and Alvis C.M.Fong, "An Effective Approach for Periodic Web Personalization", Proceedings of the IEEE/ACM International Conference on Web Intelligence. IEEE,2006.

[3] Catlegde L. and Pitkow J., "Characterising browsing behaviours in the world wide Web", Computer Networks and ISDN systems, 1995.

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

[5] Li Chaofeng, " Research and Development of Data Preprocessing in Web Usage Mining", International Conference on Management Science and Engineering , 2006.

[6] Koichiro Mihara, Masahiro Terabe and Kazuo Hashimoto, " A Novel Web Usage Mining Method Mining and Clustering of DAG Access Patterns Considering Page Browsing Time" ,Proceedings of Web Information Systems and Technologies ,2008.

[7] Mehdi Heydari, Raed Ali Helal, and Khairil Imran Ghauth, " A Graph-Based Web Usage Mining Method Considering Client Side Data", International Conference on Electrical Engineering and Informatics, IEEE, 2009.

[8] Murat Ali Bayir, Ismail Hakki Toroslu, Ahmet Cosar and Guven Fidan " Discovering more accurate Frequent Web Usage Patterns", arXiv0804.1409v1, 2008.

[9] Murat Ali Bayir, Ismail Hakki Toroslu, Ahmet Cosar and Guven Fidan " Smart Miner:A new Framework for Mining Large Scale Web Usage Data", International World Wide Web Conference Committee, ACM, 2009.

[10] Robert.Cooley,Bamshed Mobasher and Jaideep Srinivastava, "Data Preparation for Mining World Wide Web Browsing Patterns", Journal of Knowledge and Information Systems,1999.

[11] Robert.Cooley,Bamshed Mobasher, and Jaideep Srinivastava, " Web mining:Information and Pattern Discovery on the World Wide Web", In International conference on Tools with Artificial Intelligence, pages 558-567, Newport Beach, IEEE,1997.

[12] Robert F.Dell ,Pablo E.Roman, and Juan D.Velasquez, "Web User Session Reconstruction Using Integer Programming," , IEEE/ACM International Conference on Web Intelligence and Intelligent Agent,2008.

[13] Spilipoulou M.and Mobasher B, Berendt B.,"A framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis", INFORMS Journal on Computing Spring ,2003

[14] Wei Liu and Jungie Chen, "Research for Web Usage Mining Model", International Conference on Computational Intelligence for Modelling Control and Automation, IEEE,2006.

[15] Yan Li, Boqin FENG and Qinjiao MAO, "Research on Path Completion Technique in Web Usage Mining", International Symposium on Computer Science and Computational Technology, IEEE,2008.

[16] Zhang Huiying, Liang Wei," An Intelligent Algorithm of Data Pre-processing in Web Usage Mining" , Proceedings of the 5th World Congress on Intelligent Control and Automation, 2004.

7. AUTHORS PROFILE

Mrs. V. Chitraa is a doctoral student in Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu. She is working as an Assistant Professor in CMS college of Science and Commerce, Coimbatore. Her research interest lies in Database, Knowledge mining. She has presented 3 papers and published 2 papers in reputed international journals.

Dr. Antony Selvadoss Thanamani is working as Reader in NGM college, Pollachi with a teaching experience of about 22 years. His research interests includes knowledge management, web mining, networks, mobile computing, telecommunication. He has guided 41 M.Phil scholars, attended 15 conferences, presented 30 papers, published about 8 books and 16 papers.