ISSN 2278-7763

Efficient Preprocessing technique using Web log mining

Sheetal A. Raiyani[1] , Shailendra jain[2]

[1]Department of CSE(Software System),Technocrats Institute of Technology,Bhopal,India; [2]Department of CSE, Technocrats Institute of Technology,Bhopal,India.
Email: sheetal.raiyani@gmail.com[1], shailendrajain78@rediffmail.com[2]

ABSTRACT

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. Efficient 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

1 INTRODUCTION

With the rising status of the Web and the fast progress of the Web technology, hundreds of millions of communication are processed every day through the Web. Web servers keep log entries for all communication that are accessing their sites, and the sizes of those log files are increasing by tens of megabytes every day. Server logs disclose an massive amount of information about visitors, server behavior, changes in sites, and potential benefits of new technical developments. Most institutions have not been able to perform an effective use of Web server log files for enhancing and improving server performance and design we need to identify way of user accessing the web pages in particular session time. Web usage mining is the application of data mining we present new techniques for preprocessing web log data and for identifying unique users and sessions from the data. We present a fast active Distinct user identification algorithm with time complexity O(n). The algorithm uses both an IP address and a finite users' inactive time to identify different users in the web log for the session identification. In addition, we present extra cleaning steps such as removing maintenance pages, removing redundant pages, and grouping sessions with similar session lengths. The data differs in its origin as well as in its classification. Usually, it is categorized into four groups [1].

1.1 Content:

The real data in the Web pages, i.e. the data the Web page was designed to convey to the users. This usually consists of, but is not limited to, the text and graphics.

1.2 Structure:

Data which describes the organization of the content. Intra-page structure information includes the arrangement of various HTML or XML tags within a given page. The principal kinds of inter-page structure information are hyper-links connecting one page to another.

1.3 Usage:

Data that describes the pattern of usage of Web pages, such as IP addresses, page references, and the date and time of accesses. Typically, the usage data comes from an Extended Common Log Format (ECLF) server log.

1.4 User Profile:

Data that provied provides demographic information about the users of the Web site. This includes registration data and the user profile information.

2 WEB USAGE MINING

Web usage mining is the application of data mining Techniques to discover interesting usage patterns from web data, in order to understand and better serve the needs of web-based applications. It tries to make sense of the data generated by the web surfer's Sessions/behaviors. While the web content and structure mining utilize the primary data on the web, web usage mining mines the secondary data derived from the in-

ISSN 2278-7763

teractions of the users while interacting with the web. Registration data, user sessions, cookies, user queries, mouse clicks, and any other data as the results of interactions. Web usage mining method based on data cube. The approach based on data cube stresses on turning web logs into structuralized data cube which can introduce various data mining technologies [3]. Web usage mining analyzes results of user interactions with a web server, including web logs; click streams, and database transactions at a web site of a group of Related sites. Web usage mining also known as web log mining. Web usage mining process can be regarded as a three-phase process consisting:

2.1 Preprocessing/ data preparation

Web log data are preprocessed in order to clean the data moves log entries that are not needed for the mining process, data integration, identify users, sessions, and so on.

2.2 Pattern discovery

Statistical methods as well as data mining methods (path analysis, Association rule, Sequential patterns, and cluster and classification rules) are applied in order to detect interesting patterns.

2.3 Pattern analysis phase

Discovered patterns are analyzed here using OLAP tools, knowledge query management mechanism and intelligent agent to filter out the uninteresting rules/patterns.

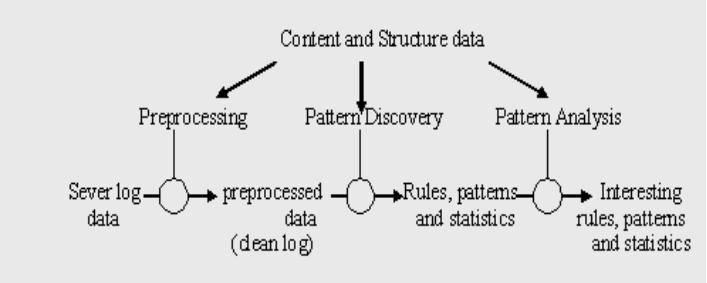

Fig 1: Process of Web Usage Mining.

3 Efficient Preprocessing Technique

The data preparation process is often the most time consuming and computationally intensive step in the Web usage mining process. The process may involve preprocessing the original data, integrating data from multiple sources, and transforming the integrated data into a form suitable for input into specific data mining operations. This process is known as data Preparation [5]. 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.

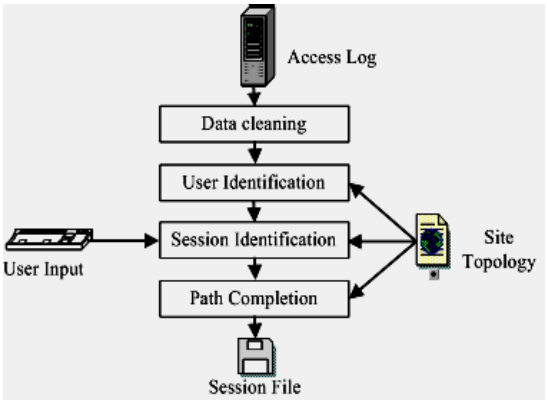

Fig-2: Complete Preprocessing Technique

3.1 Data Cleaning

The principle of data cleaning is to reduce extraneous 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, extraneous records in web access log will be eliminated during data cleaning. Since the target of Web Usage Mining is to get the user's trek patterns, following two kinds of records are unnecessary and Should 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 groping the Status field of every record in the web access log, the records with status codes over 299 or under 200 are removed. It should be pointed out that different from most other researches, records having value of POST or HEAD in the Method field are reserved in present study for acquiring more accurate referrer information.

That may not provide useful information in analysis or data mining tasks.

No	Object Type	Unique Users	Requests	Bytes In
1	*.gif	1	46	89.00 KB
2	*.js	1	37	753.95 KB
3	*.aspx	1	34	397.05 KB
4	*.png	1	31	137.67 KB
5	*.jpg	1	20	224.72 KB
6	Unknown	1	15	15.60 KB
7	*.ashx	1	15	104.79 KB
8	*.axd	1	13	274.81 KB
9	*.css	1	8	71.78 KB
10	*.dll	1	7	26.41 KB
11	*.asp	1	4	1.26 KB
12	*.html	1	3	2.17 KB
13	*.htm	1	2	69.87 KB
14	*.pli	1	2	24.92 KB

Fig -3: Example of Web Log With different Extension

ISSN 2278-7763

Algorithm: Data cleaning
The proposed algorithm for data cleaning is given below:
Input: Web server Log File
Output: Log Database
Step1: Read LogRecord from Web Server Log File
Step2: If (LogRecord .url-stem (gif.jpegjpg.cssjs))
AND (LogRecord.method='GET') AND
LogRecord.Sc-status<>(301,404,500)AND
(LogRecord.Useragent<>Crawler.Spider.Robot))
Then insert LogRecord in to LogDatabase.
End of If condition.
Step3: Repeat the above two steps until eof
(Web Server Log File)
Step4: Stop the process.

3.2 User Identification

User's identification is, to categorize 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 sequence of web pages user surf in a particular access. 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 to distinguish 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 methodsThe Web Usage Mining methods that rely on user cooperation are the easiest ways to deal with this problem. However, it's difficult because of security and privacy use the following heuristics to identify the user:

- Each IP address represents one user;
- For more logs, if the IP address is the same, but the agent log shows a change in browser software or operating system, an IP address represents a different user ;
- Using the access log in conjunction with the referrer logs and site topology to construct browsing paths for each user. If a page is requested that is not directly reachable by a hyperlink from any of the pages visited by the user, there is another user with the same IP address.

3.3 Session Identification

A user session means a delimited set of user clicks (click stream) across one or more Web servers. The goal of session identification is to divide the page accesses of each user into individual sessions. At present, the methods to identify user session include timeout mechanism and maximal forward Reference mainly. The following is the rules we use to identify user session in our experiment:
1) If there is a new user, there is a new session;
2) In one user session, if the refer page is null, there is a new session;
3) If the time between page requests exceeds a certain limit (30 mintes), it is assumed that the user is starting a new session.

3.4 Path Completion

As the reality of local cache and proxy server, there are many important accesses that are not recorded in the access log. The task of path completion is to fill in these missing page references. Methods similar to those used for user identification can be used for path completion. If a page request is made that is not directly linked to the last page a user requested, the referrer log can be checked to see what page the request came from. If the page is in the user's recent request history, the assumption is that the user backtracked with the "back" button available on most browsers, calling up cached versions of the pages until a new page was requested. If the referrer log is not clear, the site topology can be used to the same effect. If more than one page in the user's history contains a link to the requested page, it is assumed that the page closest to the previously requested page is the source of the new request. [11]

4 RELEATED WORK

User identification an important issue is how exactly the users have to be distinguished. It depends mainly on the task for the mining process is executed. In certain cases the users are identified only with their IP addresses [6]. This can provide an acceptable result for short time periods (minutes or hours) or when the expected results from the data mining task do not need more precisely information about the unique web users. For example in case of selecting frequently visited pages for server side caching, or preloading the next page of common navigational paths.
In other cases some heuristics are used for better identification of the users. In [7][6] the different methods are grouped into two classes, the one is the class of the proactive methods and the other is that of the reactive methods. Proactive strategies aim at differentiating the users before or during the page request while reactive strategies attempt to associate individuals with the log entries after the log is written. Proactive strategies can be simple user authentication with forms, using cookies or using dynamic web pages that are associated with the browser invoking them. Reactive strategies work with the recorded log files only, and the different users will be distinguished by their navigational patterns, download timing sequence or some other heuristics based on some assumption regarding their

ISSN 2278-7763

behavior. For example in [8][6] web users are distinguished based on their navigational patterns using clustering methods.

4.1 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 lots of 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 personage 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 [9], 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.

The success of the web site cannot be measured only by hits and page views. Unfortunately, web site designers and web log analyzers do not usually cooperate. This causes problems such as identification unique user's, 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.

4.2 Proposed Work

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. Proposed method shows comparison not only based on User_IP somewhere same User_IP may generate the different web users, based on path which chosen by any user and access time with referrer page we find out the distinct web user.

Definition: given a clean and filtered web log file and record set web log file

Input: Log Database
Output: Unique Users Database
Step 1: Initialize
IPLIst=0;UsersList=0;BrowserList=0;
OSList=0;No-Of-users=0;
Step 2: Read Record From LogDatabase
Step 3: If Record.IP address is not in IPList
Then add new Record>IPaddress in to IPList
Add Record.Browser in to BrowserList
Add Record.OS in to OSList
increment count of No-Of-users
insert new user in to UserList.
Else
If Record.IPaddress is present in IPList OR
Record.Browser not in BrowserList OR
Record.OS not in ORList
Then
Increment count of No-Of-users
Insert as new user in to UserList.
End of If
End of If
Step 4: Repeat the above step 2to 3
Until eof(Log Database)
Step 5: Stop the process.

Distinct User Identification (DUI)

Definition: given a clean and filtered web log file and record set web log file

Records R= {r1,r2,r3……r.n}
where n>0

Step1: input Log database RUser of N records
Step2: Distinct User identification base
Step3:RUser=P<url, ip_addr, agent, method, operating system, status,session id,time_stamp>
Step4: RUSer=<r1,r2,r3…rn> where n!=0,i=0
Step5: while(i<n)
Step6: read Logdatabase RUser
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: i=i+1;
Step10:end loop
Setp11:end

5 Result and Analysis of Experiment

To validate the effectiveness and efficiency of our methodology mentioned above, we have made an experiment with the web server log of the library of RK University rku.ac.in. The initial data source of our experiment is from JAN 1, 2012 to Aug 3, 2012, which size is 129MB. Our experiments were performed on a 2.8GHz Pentium ,CPU, 512MB of main memory, Windows 2000 professional, SQL Server 2000 and JDK 1.5. Figure is the results of our experiment. After data cleaning, the number of requests declined from 747890 to 112783.Figure shows the detail changes in data cleaning.

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

Fig.4: Experiment on Log files

ISSN 2278-7763

Monthly Statistics for March 2012		
Total Hits		12543
Total Files		9266
Total Pages		2244
Total Visits		1064
Total KBytes		1656278
Total Unique Sites		769
Total Unique URLs		258
Total Unique Referrers		193
Total Unique User Agents		218
	Avg	Max
Hits per Hour	20	154
Hits per Day	482	807
Files per Day	356	586
Pages per Day	86	192
Visits per Day	40	55
KBytes per Day	63703	315205
Hits by Response Code		
Code 200 - OK		9266
Code 206 - Partial Content		121
Code 301 - Moved Permanently		6
Code 304 - Not Modified		414
Code 404 - Not Found		2736

Fig.5: No of Pages Viewed by Users

6 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. Here 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 useful in identification method.

References

[1] Chen L, Sycara K (1998) A Personal Agent for Browsing and Searching. In Proceedings of the 2nd International Conference on Autonomous Agents, Minneapolis/St. Paul,May 9-13, pp132-139

[2] W.-K. Chen, *Linear Networks and Systems.* Belmont, Calif.: Wadsworth, pp. 123-135, 1993. (Book style)

[3] H. Poor, "A Hypertext History of Multiuser Dimensions," *MUD History,* http://www.ccs.neu.edu/home/pb/mud-history.html. 1986. (URL link *include year)

[4] K. Elissa, "An Overview of Decision Theory," unpublished. (Unplublished manuscript)

[5] R. Nicole, "The Last Word on Decision Theory," *J. Computer Vision,* submitted for publication. (Pending publication)

[6] C. J. Kaufman, Rocky Mountain Research Laboratories, Boulder, Colo., personal communication, 1992. (Personal communication)

[7] D.S. Coming and O.G. Staadt, "Velocity-Aligned Discrete Oriented Polytopes for Dynamic Collision Detection," *IEEE Trans. Visualization and Computer Graphics*, vol. 14, no. 1, pp. 1-12, Jan/Feb 2008, doi:10.1109/TVCG.2007.70405. (IEEE Transactions)

[8] S.P. Bingulac, "On the Compatibility of Adaptive Controllers," *Proc. Fourth Ann. Allerton Conf. Circuits and Systems Theory*, pp. 8-16, 1994. (Conference proceedings)

[9] H. Goto, Y. Hasegawa, and M. Tanaka, "Efficient Scheduling Focusing on the Duality of MPL Representation," *Proc. IEEE Symp. Computational Intelligence in Scheduling (SCIS '07)*, pp. 57-64, Apr. 2007, doi:10.1109/SCIS.2007.367670. (Conference proceedings)

[10] Li Chaofeng Research and Development of Data Preprocessing in Web Usage Mining, School of Management, South-Central University for Nationalities,Wuhan 430074, P.R. China

[11] E.E. Reber, R.L. Michell, and C.J. Carter, "Oxygen Absorption in the Earth's Atmosphere," Technical Report TR-0200 (420-46)-3, Aerospace Corp., Los Angeles, Calif., Nov. 1988. (Technical report with report number)

[12] L. Hubert and P. Arabie, "Comparing Partitions," *J. Classification,* vol. 2, no. 4, pp. 193-218, Apr. 1985. (Journal or magazine citation)

[13] R.J. Vidmar, "On the Use of Atmospheric Plasmas as Electromagnetic Reflectors," *IEEE Trans. Plasma Science*, vol. 21, no. 3, pp. 876-880, available at http://www.halcyon.com/pub/journals/21ps03-vidmar, Aug. 1992. (URL for Transaction, journal, or magzine)

[14] J.M.P. Martinez, R.B. Llavori, M.J.A. Cabo, and T.B. Pedersen, "Integrating Data Warehouses with Web Data: A Survey," *IEEE Trans. Knowledge and Data Eng.*, preprint, 21 Dec. 2007, doi:10.1109/TKDE.2007.190746.(PrePrint)