ISSN : 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 6, August 2012

Advanced Preprocessing using Distinct User Identification in web log usage data

Sheetal A. Raiyani[1], Shailendra Jain[2], Ashwin G. Raiyani[3]

Department of CSE (Software System), Technocrats Institute of Technology, Bhopal, India[1]

Department of CSE, Technocrats Institute of Technology, Bhopal, India[2]

Department of CE, RK University, Gujarat, India[3]

Abstract—*Millions of visitors interact daily with web sites around the world. Huge amount of data are being generated and these information could be very prized to the company in the field of accepting Customer's behaviors. In this paper a complete preprocessing methodology having data cleaning, Enhanced preprocessing technique one of the User Identification which is key issue in preprocessing technique phase is to identify the web users. Traditional User Identification is based on the site structure by using some heuristic rules. In most cases relationship between pages are based on the site topology which reduced the efficiency of identification solve this problem we introduced proposed Technique DUI (Distinct User Identification) based on IP address ,Agent ,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 frequent access behavior get better the overall designing and performance of future access. Experiments have proved that advanced data preprocessing technique can enhanced the quality of data preprocessing results.*

Keywords— **Web usage mining, Preprocessing, User identification, Session time, Server log**

I. Introduction

Web mining refers to the use of data mining techniques to automatically retrieves, extract and analyze information for knowledge discovery from web documents and services. Web Usage Mining is a heavily researched area in the field of data mining. 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. 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. Finally, web usage mining, also known as Web Log Mining, is the process of extracting interesting Patterns in web access logs.

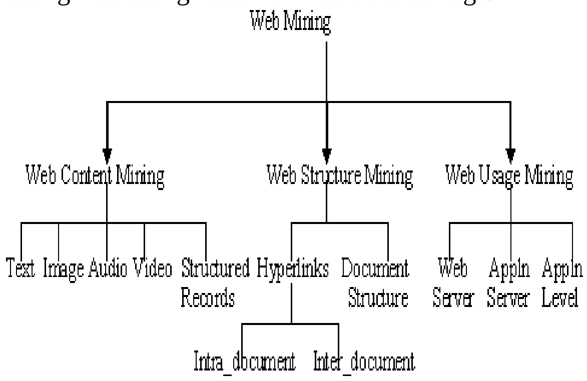


Fig-1 Taxonomy of Web mining

II. 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 interactions 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:

- Preprocessing/ data preparation - web log data are preprocessed in order to clean the data – removes log entries that are not needed for the mining process, data integration, identify users, sessions, and so on
- 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.
- 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.

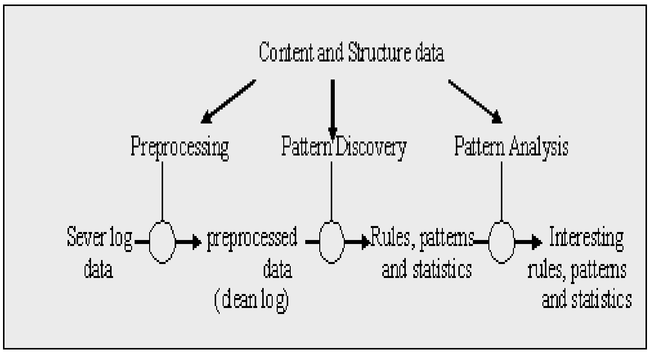

Content and Structure data
Preprocessing
Pattern Discovery
Pattern Analysis
Sever log data
preprocessed data (clean log)
Rules, patterns and statistics
Interesting rules, patterns and statistics

Fig-2: Web Usage Mining

After discovering patterns from usage data, a further analysis has to be conducted. The most common ways of analyzing such patterns are either by using query or by loading the results into a data cube and then performing OLAP operations[3]. Then, visualization techniques are used for a results interpretation. The discovered rules and patterns can then be used for improving the system performance / for making modifications to the web site. The purpose of web usage mining is to apply statistical and data mining techniques to the preprocessed web log data, in order to discover useful patterns. Usage mining tools discover and predict user behavior in order to help the designer to improve the web site, to attract visitors, or to give regular users a personalized and adaptive service. The applications are Extract statistical information and discover interesting user patterns, Cluster the user into groups according to their navigational behavior, Discover potential correlations between web pages and user groups, Identification of potential customers for ecommerce ,Enhance the quality and delivery of Internet information services to the end user ,Improve web server system performance and site design, Facilitate personalization.

III. WEB LOG FORMAT

The web usage data includes the data from web server logs, proxy server logs, browser logs, and user profiles. (The usage data can also be split into 3 different kinds on the basis of the source of its collection: on the server side (there is an aggregate picture of the usage of a service by all users), the client side (while on the client side there is complete picture of usage of all services by a particular client), and the proxy side (with the proxy side being somewhere in the middle). Web Server logs are plain text (ASCII) files, that is Independent from the server platform. There are some Distinctions between server software, but traditionally there are four types of server logs.

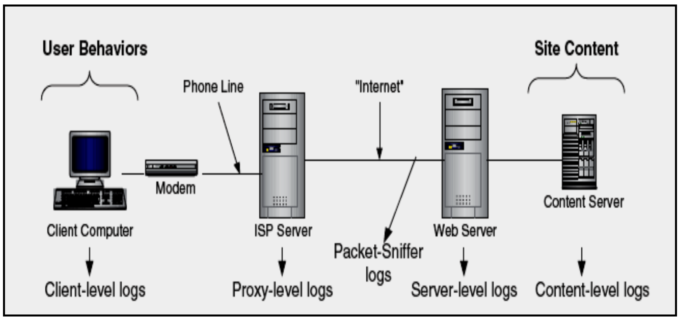


Fig-3 Different types of log

Currently, there are three formats available to record log files:-W3C Extended Log file Format-Microsoft IIS Log File-NCSA Common Log file Format.

The W3C Extended log file format, Microsoft IIS log file format, and NCSA log file format are all ASCII text formats. The W3C Extended and NCSA formats record logging data in four-digit year format. The Microsoft IIS format uses a two digit year format for years 1999 and earlier and a four-digit format thereafter. The Microsoft IIS log format is provided for backward compatibility with earlier IIS versions[2]. A web server log file contains requests made to the web server, recorded in chronological order. The most popular log file formats are the Common Log Format (CLF) and the extended CLF. A common log format file is created by the web server to keep track of the requests that occur on a web site. A standard log file has the following format

<ip_addr><base_url><date><method><file><Protoc-ol><code><bytes><referrer><user_agent>

Fig-4: CLF Log Format

IV. 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.

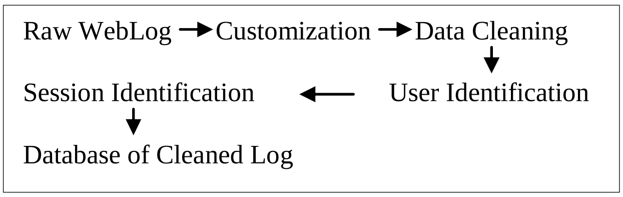


Fig-5 : Preprocessing Technique

A. 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 [1], 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 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 examining the Status field of every record in the web access log, the records with status codes over 299 or under 200 are removed.

B. 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 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 methods, this step will be discussed later in detail.

C. 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 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[4]. while 30 minutes is the default timeout. 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.[4]

The session identified by 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.

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

A. 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.

B. Proposed Method Distinct User Identification (DUI)

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

```
    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;
```

ISSN : 2278 – 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 6, August 2012

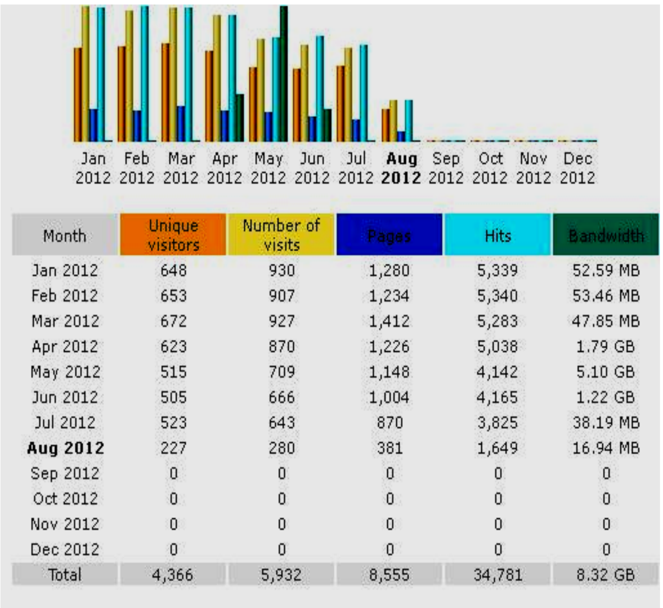
Step10:end loop
Setp11:end

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

VI. 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 Ⅳ 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.

Fig-6: Result of Experiment



Month	Unique visitors	Number of visits	Pages	Hits	Bandwidth
Jan 2012	648	930	1,280	5,339	52.59 MB
Feb 2012	653	907	1,234	5,340	53.46 MB
Mar 2012	672	927	1,412	5,283	47.85 MB
Apr 2012	623	870	1,226	5,038	1.79 GB
May 2012	515	709	1,148	4,142	5.10 GB
Jun 2012	505	666	1,004	4,165	1.22 GB
Jul 2012	523	643	870	3,825	38.19 MB
Aug 2012	227	280	381	1,649	16.94 MB
Sep 2012	0	0	0	0	0
Oct 2012	0	0	0	0	0
Nov 2012	0	0	0	0	0
Dec 2012	0	0	0	0	0
Total	4,366	5,932	8,555	34,781	8.32 GB

Fig-7 : Result of Proposed method DUI Experiment

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

[2] Mohd Helmy Abd,Mohd Norzali, Data Preprocessing on Web Server log for Generalized Association Rule Mining. World Academy of Science, Engineering and technology,48 2008

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

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

[5] Mr. Sanjay Bapu Thakare, 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, 848-851

[6] Renáta Iváncsy, and Sándor Juhász, Analysis of Web User Identification Methods, World Academy of Science, Engineering and Technology 34 2007

[7] M. Spiliopoulou and B. Mobasher and B. Berendt and M. Nakagawa, Framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis, INFORMS Journal on Computing, 15, 2003

[8] T. Morzy, M. Wojciechowski, and M. Zakrzewicz. Web users clustering. International Symposium on Computer and Information Sciences 2000

[9] 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