A Step up in Data Cleaning and User identification of Preprocessing on Web Usage data

Manisha Valera, Shaktisinh Parmar

Abstract— **Preprocessing is an important Step in Web Usage Mining. Web Log files are the great source of knowledge which can be used to analyze behavior of online users. Data Cleaning, User identification and Session Identification are involved in Data Preprocessing. The time spent by user on the web page is considered for calculating session. There are so many pattern mining methods which can be applied on preprocessed data. The preprocessing techniques will maximize the quality of pattern mining methodologies and the results can be used for recommender systems to find the behavior of a user.**

***Index Terms*—Preprocessing, Session Identification, User Identification, Web Usage Mining.**

I. INTRODUCTION

The application of data mining techniques is to discover interesting usage patterns from Web data, in order to recognize and better serve the requirements of Web-based applications. Usage data holds the identity of Web users along with their browsing behavior at a Web site. Basically there are three main tasks for performing Web Usage Analysis.

Web mining can be categorized into three areas of interest based on which part of the web to mine [1]:

A. Web Content Mining

It deals with discovering significant and useful facts from web page contents. It contains unstructured information like text, image, audio, and video.

B. Web Structure Mining

It deals with discovering and modeling the link structure of the web. Web structure mining aims to generate structural abstract about web sites and web pages.

C. Web Usage Mining

It is the application of data mining techniques to discover interesting usage patterns from Web data, in order to know and better serve the needs of Web-based applications.

Manuscript received Dec, 2014.

***Manisha Valera**, Computer Engineering, Indus University, Ahmedabad, Gujarat , India*

***Shaktisinh Parmar**, Computer Engineering, C U Shah University, Surendranagar, Gujarat,Iindia*

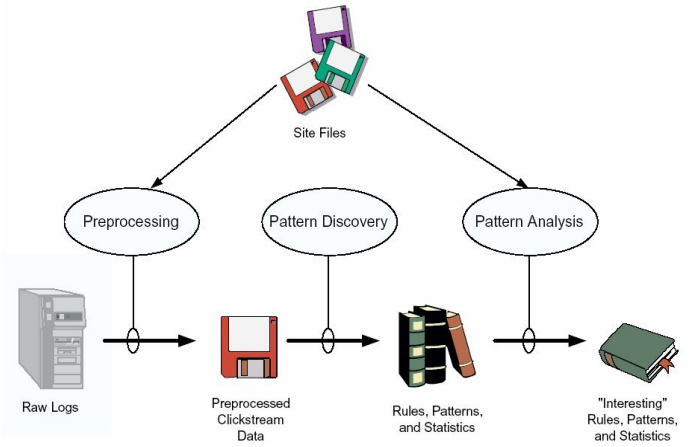

Fig.1 Process of Web Usage Mining

Six major steps followed in web usage mining are [8]:

1) Data collection Web log files, which keeps track of visits of all the visitors.
2) Data Integration Integrate multiple log files into a single file
3) Data preprocessing Cleaning and structuring data to prepare for pattern extraction
4) Pattern extraction Extracting interesting patterns
5) Pattern analysis and visualization Analyze the extracted pattern
6) Pattern applications Apply the pattern in real world problems

II. WEB USAGE MINING PROCESS

The main processes in Web Usage Mining are:

A. Preprocessing

Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Generally used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more simply and well processed for the purpose of the user.

ISSN: 2278 – 1323

B. Pattern Discovery

Web Usage mining can be used to reveal patterns in server logs but is often carried out only on samples of data. The mining process will be futile if the samples are not a good representation of the larger body of data. The following are the pattern discovery methods.

1) Statistical Analysis
2) Association Rules
3) Clustering
4) Classification
5) Sequential Patterns

C. Pattern Analysis

This is the last step in the Web Usage Mining process. After the preprocessing and pattern discovery, the obtained usage patterns are analyzed to filter uninteresting information and extract the useful information. The methods like SQL (Structured Query Language) processing and OLAP (Online Analytical Processing) can be used.

III. LITERATURE SURVEY

As presented in [4], it uses heuristic method to resolve the dilemma, 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.

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 recreate the user sessions, in this case heuristics can be used as describe. If all of the IP address, browsers and operating systems are identical, 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 before, 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 [5]. 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.

The proposed method [6] tells user behavior and it creates user cluster and site cluster. Also it gives the information like what sites are the most and least popular, which website is most commonly used by visitors and from what search engine are visitors coming frequently. In this method, if IP address is unique then similar user cluster is created; If IP address is same and user name is not unique, agent log, operating system and browser are different then distinguish user cluster is created.

In the proposed scheme [7] we can combine use of user session heuristics for web server log. One is the time oriented and the other is navigation oriented. Two navigation methods are used for user session identification. The first one is the maximum forward reference length and the second one is the reference length model to identify the user session identification. Maximum forward reference method has an advantage over the reference length model that has no require input parameter. With the use of cut off time we get more accurate user behavior which will be very helpful in finding user navigation patterns. The comparing experiment with traditional timeout algorithm and its improvements shows that the new algorithm improves the accuracy of data preprocessing.

Traditional user identification is carried out according these rules [9]:
1) Different IP address refers to different users.
2) The same IP with different operating system or different browser should be considering as different user.
3) While the IP, operating system and browsers are all the same, new user can be determined whether the requesting page can be reached by accessed pages before according to the topology of the site.

In Session Identification, Web log mining covers a long time periods, therefore users may access the site more than once. Session identification is in order to divide the access records into several accessing sequences, in which the pages are requested at the same time. Traditional session identification algorithm is based on a uniform and fixed timeout. While the interval between two sequential requests exceeds the timeout, new session is determined.

IV. PREPROCESSING

The idea of data preprocessing is to extract useful information from raw web log and then transform these data in to the form necessary for pattern discovery. Due to large amount of irrelevant information in the web log, the original log cannot be directly used in the web log mining procedure, hence in data preprocessing phase, raw Web logs have to be cleaned, analyzed and converted for further step. The input for the proposed system is a web server log data and it comprises IP address, access time, HTTP request method used, URL of the referring page and browser name. It is difficult for these web server log data to be directly used to mine the desired sequential pattern mining process [2]. So, due to that phenomenon, the following preprocessing techniques need to be used in the raw web server log data.

The log entry contains various fields which need to be separate out for the processing. The process of separating field from the single line of the log file is known as field extraction. The server used different characters which work as separators. The most used separator character is or 'space' character. The Data cleaning algorithm is given below.

The process of data cleaning is removal of outliers or irrelevant data. The Web Log file is in text format then it is required to convert the file in database format and then clean the file. First, all the fields which are not required are removed and finally we will have the fields like date, time, client ip, URL access, Referrer and Browser used/ Access log files consist of large amounts of HTTP server information. Analyzing, this information is very slow and inefficient without an initial cleaning task. All log entries with file name suffixes such as gif, JPEG, jpeg, GIF, jpg, JPG can be eliminated since they are irrelevant [10].Web robot (WR) (also called spider or bot) is a software tool that periodically a web site to extract its content[11]. To identify web robot requests the data cleaning module removes the records containing "Robots.txt" in the requested resource name (URL). The HTTP status code is then considered in the next process of cleaning by examining the status field of every record in the web access log, the records with status code except 200 are removed because the records with status code 200, gives successful response[12].

Data Cleaning:
Input: Log File
Output: cleaned Log Table
Algorithm:

1) Open a DB connection
2) Create a table to store log data
3) Open Log File
4) Read all fields contain in Log File
5) for each record in Log Table
   if (status code='200' and Method='Get')
   {
      Remove fields where (CS-URI-STEM = '.jpg' or '.jpeg' or '.gif' or '.png' or 'robot.txt' or '.css')
   }
Next Record
End

User Identification:
Input: N records of web log file
Output: User sets identified
Algorithm:
1) Repeat steps
   if(ip address of first log entry == ip address of second log entry)
   {
      Compare the user agent of both entries
      If both user agents are same
      Identify both entries are from same user.

   }
2) else assume as different users.
   Until last entry

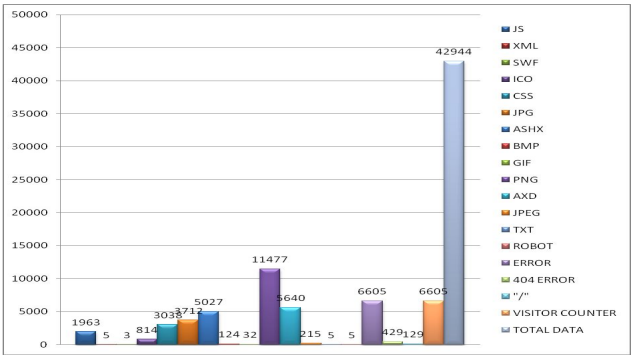

Fig. 2 Preprocessing of Web Log Data

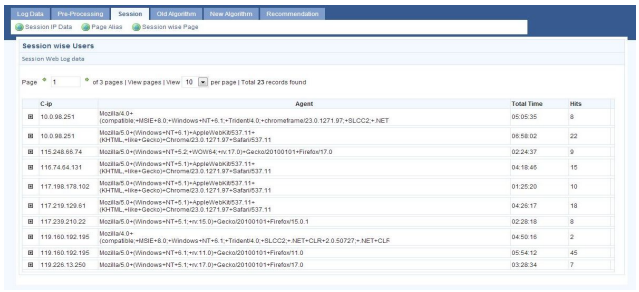

Fig. 3 Screenshot of User Identification

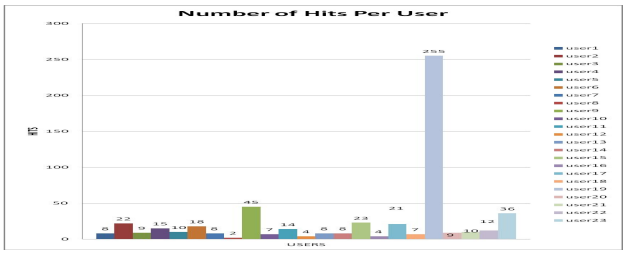

Fig. 4 Numbers of Hits per User

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 effectiveness and the accuracy of user identification without usage of site topology.

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.
BT[i] = (BT[i+1]-BT[i]) – Time-Taken[i]

ISSN: 2278 – 1323

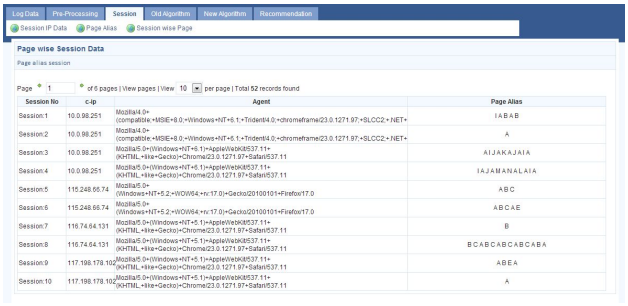

Fig. 5 Screenshot of Session's Web access Sequence

The different sessions belonging to different users should be identified.. A session is a group of activities performed by a user when he is navigating through a given site. For web server logs, all users' access activities of a website are recorded by the Web server of the website. Each user access record contains the client IP address, request time, requested URL, HTTP status code, etc. Web logs can be regarded as a collection of sequences of access events from one user or session in timestamp ascending order. Here we are defining different sessions according to their time stamp order, with a time interval of 1 hour for each session. For example, the Session S1 includes all the web pages accessed during time duration of 09:30-10:30 that are P1,P2,P1,P3,P4, similarly S2 has P1,P3,P5 page sequence and S3 has P4,P1 page sequence[3].

V. CONCLUSION

In this Research we have presented data cleaning and distinct user identification technique which enhance the pre-processing steps of web log usage data in data mining. Using user identification we find out distinct user based on their attended session time. 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 can be performed on preprocessed data.

REFERENCES

[1] Udayasri.B, Sushmitha.N, Padmavathi.S, "A LimeLight on the Emerging Trends of Web Mining" , *Special Issue of International Journal of Computer Science & Informatics (IJCSI),* ISSN (PRINT):2231–5292,Vol.-II,Issue-1,2

[2] Utpala Niranjan, Dr.R.B.V. Subramanya, Dr.V.Khanaa,"An Efficient System Based On Closed Sequential Patterns for Web Recommendations", *International Journal of Computer Science Issues,* Vol. 7, Issue 3, No 4, May 2010.

[3] Dheeraj Kumar Singh, Varsha Sharma, Sanjeev Sharma "Graph based Approach for Mining Frequent Sequential Access Patterns of Web pages", *International Journal of Computer Applications* (0975 – 8887) Volume 40– No.10, February 2012.

[4] 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

[5] 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

[6] S. Umamaheswari, S. K. Srivatsa,"Algorithm for Tracing Visitors' On-Line Behaviors for Effective Web Usage Mining",*International Journal of Computer Applications, (*0975 – 8887) Volume 87 – No.3, February 2014

[7] Priyanka Patel, Mitixa Parmar," Improve Heuristics for User Session Identification through Web Server Log in Web Usage Mining", *International Journal of Computer Science and Information Technologies*, Vol. 5 (3) , 2014, 3562-3565

[8] Manisha Valera, Kirit Rathod ," A Novel Approach of Mining Frequent Sequential Pattern from Customized Web Log Preprocessing", *International Journal of Engineering Research and Applications (IJERA),* ISSN: 2248-9622 ,Vol. 3, Issue 1, January -February 2013, pp.269-380

[9] He Xinhua, Wang Qiong," Dynamic Timeout-Based a Session Identification Algorithm" *Electric Information and Control Engineering (ICEICE), 2011 International Conference* ,ISBN 978-1-4244-8036-4, 15-17 April 2011.

[10] Navin Kumar Tyagi, A.K. Solanki & Sanjay Tyagi. "An Algorithmic approach to data preprocessing in Web usage mining*", International Journal of Information Technology and Knowledge Management* , July-December 2010, Volume 2, No. 2, pp. 279-283 .

[11] J. Vellingiri and S. Chenthur Pandian, "A Novel Technique for Web Log mining with Better Data Cleaning and Transaction Identification", *Journal of Computer Science7 (5): 683-689*,2011 ISSN 1549-3636 © 2011 Science Publications

[12] Priyanka Patil and Ujwala Patil, " Preprocessing of web server log file for web mining*", National Conference on Emerging Trends in Computer Technology (NCETCT-2012*), April 21, 2012

Manisha Valera has completed M.E in Computer Engineering from GTU in 2013. And right now she is working as an Assistant Professor in Indus University. Her areas of interest are Web mining, data mining, Big data. She has 4 research publications, in those 3 are of international journal and 1 is of international conference.

Shaktisinh S. Parmar has completed M.S. in Computer Science. And right now working as an Assistant Professor in C U Shah University. His areas of interest are Web mining, Software Engineering, Computer Algorithms. he has 4 research publications, in those 3 are of national journal and 1 is of international conference.

