ISSN:0975-9646

Improve Heuristics for User Session Identification through Web Server Log in Web Usage Mining

Priyanka Patel[1], Mitixa Parmar[2]

[1]*IT Department, Parul Institute of Engineering and Technology/GTU,India,*
[2]*Faculty of IT Department, Parul Institute of Engineering and Technology/GTU,India*

Abstract -In the web servers, log repositories plays a key role as it keeps record of user pattern for different users and thus it is great source of knowledge. Web Usage Pattern is process of getting the web user browsing patterns by analyzing their navigational behavior. A modern process of identifying a web user session is the average time spent by user on pages. The time spent by user on web site is taken with high importance in this paper. To get the clusters of session we applied rough set clustering method. With the combination of statistical result and importance degree of page the initial time out is calculated. After that for user session identification session timeout is dynamically adjusted. Fixed value of session time may not give accurate user sessions and so here average user time spent by user on web site is considered. Web user logs contain all details about user behavior and accurate user session will give more accurate user navigation behavior or pattern.

I. INTRODUCTION

There are the three subpart of the data mining.1)Web structure mining 2) Web content mining 3) web usage mining. Web usage mining is the third category in web mining. This type of web mining allows for the collection of Web access information for Web pages. This usage data provides the paths leading to accessed Web pages. This information is often gathered automatically into access logs via the Web server. CGI scripts offer other useful information such as referrer logs, user subscription information and survey logs. This category is important to the overall use of data mining for companies and their internet/ intranet based ap plications and information access.

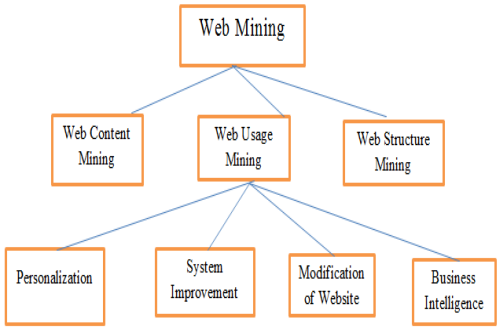

Figure 1.1 Web Mining Categories

Usage mining allows companies to produce productive information pertaining to the future of their business function ability. Some of this information can be derived from the collective information of lifetime user value, product cross marketing strategies and promotional campaign effectiveness. The usage data that is gathered provides the companies with the ability to produce results more effective to their businesses and increasing of sales. Usage data can also be useful for developing marketing skills that will out-sell the competitors and promote the company's services or product on a higher level. Usage mining is valuable not only to businesses using online marketing, but also to e-businesses whose business is based solely on the traffic provided through search engines. The use of this type of web mining helps to gather the important information from customers visiting the site. This enables an in-depth log to complete analysis of a company's productivity flow. E-businesses depend on this information to direct the company to the most effective Web server for promotion of their product or service.

The first is usage processing, used to complete pattern discovery. This first use is also the most difficult because only bits of information like IP addresses, user information, and site clicks are available. With this minimal amount of information available, it is harder to track the user through a site, being that it does not follow the user throughout the pages of the site. The second use is content processing, consisting of the conversion of Web information like text, images, scripts and others into useful forms. This helps with the clustering and categorization of Web page information based on the titles, specific content and images available.

Finally, the third use is structure processing. This consists of analysis of the structure of each page contained in a Web site. This structure process can prove to be difficult if resulting in a new structure having to be performed for each page.

The Procedure of data preprocessing.1) Data cleaning 2) user identification 3)Session identification 4)Path completion 5) Transaction identification

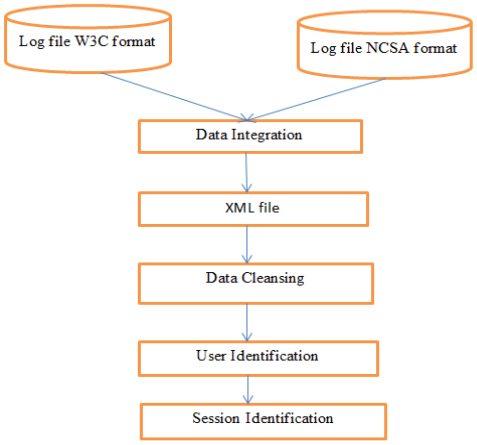

Figure 2.1 Log file Preprocessing

a) Data cleaning[1]

Data cleaning means deleting needless data, which can't reflect the characters of user's accessing behavior. It covers with some aspects as follows 1) Irrespective attribute: The attributes paid attention to include users IP address, URL pages requested, and accessing time and so. And other attributes should be got rid of.2) Content of pages as picture, video and audio resources and the logical units as script CSS files.3) the actions that not request for pages and fail requested pages. The actions that not request for pages failed requested pages. The former can be judged by the code of POST or GET and the later can be judged by the state code.

b) User Identification[1]

Traditional user identification is carried out according these rules:1) Different IP address refer to different users.2) The same IP with different operating system or different browser should be consider 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.

c) Session identification[1]

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.

d) Path Complement[1]

Because of local buffer and agent service buffer exists, not all requested pages are recorded in Web log. If a requested page that can be reachable by a hyperlink from any of the pages visited by the user before, we assume that the page is requested from local buffer, which should be added in the user sessions. This procedure is called path complement. Path complement can generally be accomplished according to the topology of site and the user's accessing sequence.

e) Transaction identification[1]

For some specific mining algorithm, user session still has a large scale, which needed to be divided for a smaller scale. Transaction identification means finding out significant accessing sequences, which also called accessing transaction. In other words, transaction identification is a procedure of grouping user session. The former uses the user's accessing behavior to divide transaction. Once the user step back to visit the pages that visited before, the boundary of transaction is determined. The later divides the pages into content pages and assistant pages according to the accessing time taken by user. While the pages turn into content pages from assistant pages, the boundary is determined.

II. LITERATURE SURVEY

In this paper to improve the accuracy of data preprocessing in web log mining, basic procedure of data preprocessing is introduces first. Then the traditional session identification algorithm is analyzed on the basis which, a session identification algorithm based on dynamic timeout is presented. The initial timeout is compared for each page according to the statistical result, combining with the importance degree of page, procedure of session identification, time out is dynamically adjusted, user sessions is determined judging by the dynamic timeout. Comparing experiment shows that the algorithm proposed can obtain a better performance on session identification[2].Only proper data preprocessing can attain right data that shows the intension of web users correctly ,and ensure the right direction of data mining. Traditional data processing contains five steps as data cleaning, user identification, session identification, path complement, transaction identification. Then a session identification algorithm based on dynamic timeout is presented.[1]

In this paper the data preprocessing, sessions identification is important step. Algorithms used so far to identify sessions use some fixed values to specify the end of a session and to mark the beginning of another. In paper we explain why the use of fixed values cause errors in identifying sessions and we propose a new method for identifying sessions based on average time of visiting web pages We implemented in Java programming language by using Net Beans IDE, two algorithms to identify sessions. The first uses a fixed value of 30 minutes (1800 seconds) to indicate the end of a session and the second using the average time spent on the pages of the website by users.[2]

In this paper author present Web Usage Mining (WUM) is the discovery of interesting knowledge from web server log. The access log files of a Web server contain a lot of details about users' on-site behavior. The validity of WUM depends on the accurate identification of user sessions implicitly recorded in these logs. This is generally difficult because of several additional factors, such as Web caching, the existence of proxy servers and the stateless service model of the HTTP protocol. Several heuristics exist to address these problems[9]. This paper reports on an investigation into the performance of a composite heuristic based on three published heuristics found in literature to identify sessions from the Web logs. We use the logs of a university Web server that records user ids for administrative reasons, which allows us to evaluate the heuristics against the concrete knowledge of user sessions. The paper also proposes a strategy for future log analyses.[3]

In this paper authors present 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.[7]

III. PROPOSED SYSTEM

In proposed scheme 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 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. The maximum forward reference transaction identification based on the time spent on a web page. Each transaction is defined as a set of web page in the path from the first page to the user session up to backward reference is made. Backward reference is the defined as web page that is already contain in the set of web pages for the current transaction. A new session is started when the next forward reference is made. This method has a maximum forward reference that web page is content page and the other pages are the navigation page. In that other method is based on assumption of time spent on a particular page classified as a content page or a navigation page. We define the assumption about the portion of navigation of surveyed web log. We can define the cut-off time that separate the web pages either content page or the navigation page. When the cut-off time known the session can be created and compare the time of a particular web page visit to the cut-off time. User spent more time than the cut-off time then the page identify as a content page and user spent less time compare to the cut-off time that page will identify as a navigation page. The cleaning process will be apply on the web log data and filter or remove noisy data like css, java script and picture file. The web log data contains the useful data to further preceed.

The cut-off time will be used to identify the navigation or a content page. To improve efficiency of user pattern analysis and to get more efficient user behavior and user session identification based on the navigation of user through the website . Every session logs have records of page requests generated by user. Length Model uses these requests and it generate length of reference from two consecutive same requests made by user. As current transaction will not have next request time it assumes all last references are end of session and ignores them in further calculations. This system has focused on cut off time calculation based on the Length request generated by Length Model. With the use of this cut off time we get more accurate user behavior which will be very helpful in finding user navigation patterns. This scheme proposed for to identify the correct and accurate session from the web log.

Pseudo code of Propose system

The pseudo code of propose algorithm will explain how the algorithm will work step by step.

Step 1: Apply Data cleaning Algorithm on Web log file

Remove noisy and unwanted entry , jpg file, css file and java script.

Step 2 : Set Initial Timeout

Initial timeout(IP)= 30 min and then the using initial timeout to identify the number of session.

Step 3: Calculate Cutoff time

```
For each IP IP_list repeat
Get CutOff[IP];
Difference = 0;
while (Difference < CutOff)
Difference = Time_Spent[NextPage] Time_Spent[CurrentPage];
If (Difference >= CutOff)
Create New Session
CutOff = (Cutoff+Difference)/2;
```

Step 4: Identify Content page or Navigation page

```
For each IP IP_list repeat
Get CutOff[IP];
While (Time_Spent[Current_Page] < cutoff)
Nevigation_Page=Current_Page;
Continue;
Content_Page=Current_Page;
Creat New Session;
```

IV .RESULT ANALYSIS

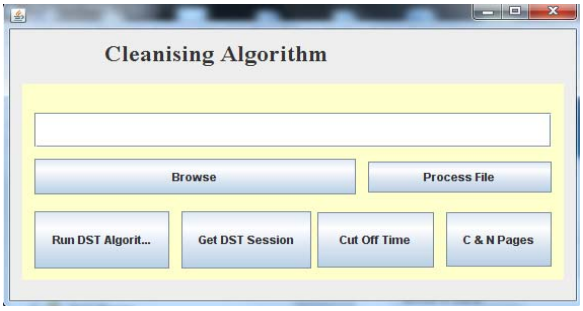

Figure 5.1 GUI of Proposed system

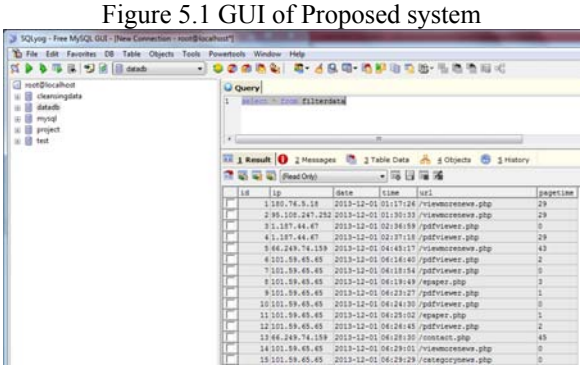

Figure 5.2 Result of Data Cleaning Algorithm

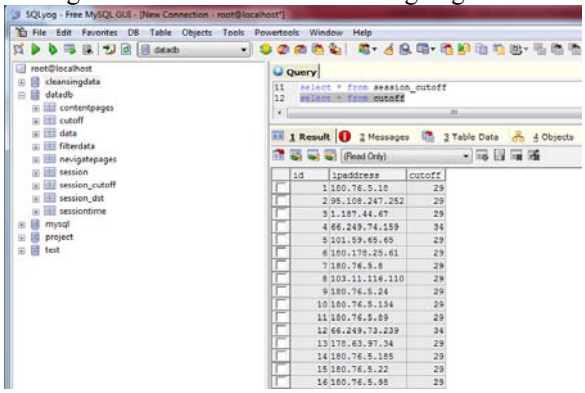

Figure 5.3 Calculate cutoff time for cleaned data

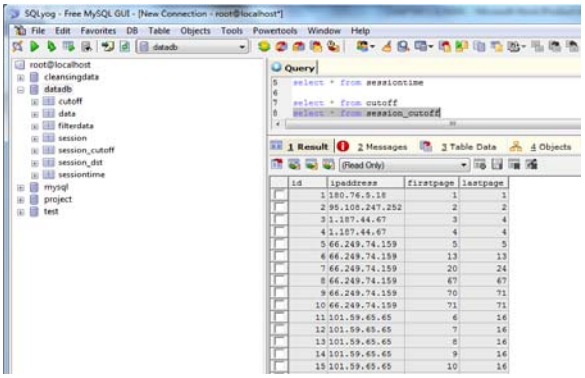

Figure : 5.4 Session Identify by Cutoff Time

Table 5.1 Comparison of Existing and Proposed method Results

	No. of session identification
Total Records	13424
After Cleaning Process	2805
Traditional Timeout Algorithm	558
Dynamic Timeout Algorithm	615
Cutoff Time Algorithm	2008

The above table shows the total number of records using during the data cleaning process than after the shows the number records after the data cleaning process. And also shows the number of session identified by traditional timeout algorithm.In dynamic adjustment of timeout calculated first the difference between the current page and the next page how much time spent on that page. After find the average of initial time set for the particular page and the difference of current page and the next visited page. In Cut off time calculated first set the cut off time for the current page and then find the new cut off time will be the average of the last cutoff time and time spent on current content page.

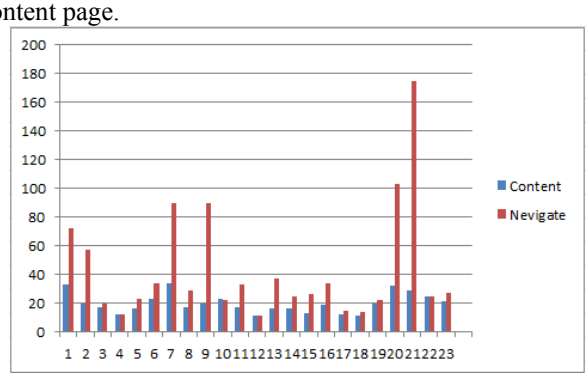

Figure 5.5 Identification of Content and Navigation Page

V. CONCLUSION

A Session identification algorithm based on dynamic timeout and another algorithm is basis on the traditional timeout algorithm, which combined with accessing habit of web users and the characters of different pages. It can also be calculate the Average and Mean time to identify the user session. If the data will use the preprocessed than the identification of session will be accurate. The comparing experiment with traditional timeout algorithm and its improvements shows that the new algorithm proposed for better performance on session identification, which improves the accuracy of data preprocessing. The comparison of the average time and all other method to use the identification to accuracy of the session and data. The modified algorithm average time depends on the web page, so we can say that using this algorithm to separate sessions better reality than using single constant values for the algorithm.

REFERENCES

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

[2] C. E. Dinuca, D. Ciobanu," Improving the session identification using the mean time" INTERNATIONAL JOURNAL OF MATHEMATICAL MODELS AND METHODS IN APPLIED SCIENCES, Volume 6, 2012.

[3] Amithalal Caldera and Yogesh Deshpande," Evaluation of Session Identification Heuristics in Web Usage Mining" Towards an ICT enabled Society,2011.

[4] Dumitru Ciobanu , Claudia Elena Dinucă," A new method for session identification in clickstream analysis" Recent Researches in Tourism and Economic Development ISBN: 978-1-61804-043-5 476 Vol.12 ,2011

[5] Dilip Singh Sisodia, Shrish Verma," Web Usage Pattern Analysis Through Web Logs: A Review" Ninth International Conference of Computer science and Software engineering Vol 2 December 2012

[6] Peng Zhu, Ming-sheng Zhao "Session Identification Algorithm for Web Log Mining" ,IEEE,2011

[7] 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

[8] Zheng Lu, Student Member, IEEE, Hongyuan Zha, Xiaokang Yang, Senior Member, IEEE, Weiyao Lin, Member, IEEE, and Zhaohui Zheng," A New Algorithm for Inferring User Search Goals with Feedback Sessions", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 3, MARCH 2013".

[9] Ms. Jyoti , Dr. A. K. Sharma, Dr. Amit Goel" A Novel Approach for clustering web user sessions using RST " International Journal on Computer Science and Engineering Vol.2(1), 2009, 56-61, ISSN : 0975-3397

[10] Lucchese, C., Orlando, S., Perego, R., Silvestri, F., & Tolomei, G. (2011, February). Identifying task-based sessions in search engine query logs. In Proceedings of the fourth ACM international conference on Web search and data mining (pp. 277-286). ACM.

[11] El-Ramly, M. and Strulia, E., Web-usage Mining and Run-time URL Recommendation for Focused Web Sites: A Case Study. "Journal of Software Maintenance and Evolution: Research and Practice". **00**: p. 1-7. John Wiley & Son, Ltd 2000

[12] Felciano, R.M. and Altman, R.B., Lamprey: Tracking Users on the World Wide Web. "AMIA Annual Fall Symposium". Hanley & Belfus. Washington, D.C. 1996

[13] Tak Woon Yan, Matthew Jacobsen, Hector Garcia-Molina, and Dayal, U., From User Access Pattern to Dynamic Hypertext Linking. "Fifth International World Wide Web Conference". Paris, France May 6-10, 1996

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