ISSN:0975-9646

A Survey: User Identification Techniques for Customization of Web Page

Sohit Shukla[1], Supriya Mishra [2], Bhawesh Kumar Thakur [3]

[1]*Rajkiya Engineering College, Ambedkarnagar, India*
[2]*Rajkiya Engineering College, Ambedkarnagar, India*
[3]*Ambalika Institute of Management & Technology, Lucknow, India*

***Abstract*—The large amount of data available online which is increasing day by day, the World Wide Web has becoming one of the most valuable resources for information retrievals and knowledge discoveries. Web mining technologies are the right solutions for knowledge discovery on the Web. Millions of visitors interact daily with web sites around the world. Huge amount of data are being generated and these information could be very valued to the company in the field of understanding Customer's behaviors. In this paper we have discussed several existing user identification techniques with their comparative study.**

***Keywords*—**User Identification, Log files, Web Usage Mining.

I. INTRODUCTION

The web is vast, varied and dynamic and thus raises the scalability, multimedia data and temporal issues respectively. The expansion of the web has resulted in a large quantity of data that is now freely available for user access. The different types of data have to be managed and organized in such a way that they can be accessed by different users effectively and efficiently.

In our study we will give emphasize on Web usage mining. As Now a day the web is not the place where only transaction has occurred. Millions of visitors interact with the web in daily life which generates an enormous amount of data. Web usage mining helps to know information about users' behaviors and their usage patterns, can lead to interesting results that go over. Web Usage Mining (WUM) is the automatic discovery of user access pattern from web servers. Organizations collect large volumes of data in their daily operations, generated automatically by web servers and collected in server access logs. It can also
provide information on how to restructure a website to service effectively [1].

Web Usage Mining (WUM) is the process of extracting knowledge from Web users access data by exploiting Data Mining technologies. It can be used for different purposes such as personalization, system improvement and site modification. Web usage mining tries to make sense of the data generated by the web surfer's session or behaviors. While the web content and structure mining utilize the real or primary data on the web, web usage mining mines the secondary data derived from the interactions of the users while interacting with the web.

Log file of the web server is a source of anonymous data about the user. The web usage data includes the
data from web server logs, proxy server logs, browser logs, and user profiles. These anonymous data represent also the problem of unique identification of the web site visitor. If we want to analyze the users' behaviour on our web site, it is not necessary to know
the identity of each visitor, but, it is very important for us to distinguish the web site visitors. The task of user identification is, to identify who access web site and which pages are accessed [4].

The analysis of Web usage does not require knowledge about a user's identity. However, it is
Necessary to distinguish among different users. Since a user may visit a site more than once, the server logs record multiple sessions for each user. The phrase user activity record is used to refer to the sequence of logged activities belonging to the same user.

In this paper we have discussed various existing user and session identification to 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 through web log data.

II. RELATED WORK

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.

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.[7]

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.

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.

Different User Identification Methods through web Log Data[2]:

1. Proactive mechanisms
2. Reactive heuristics
3. IP Address Based Identification
4. Cookie-based user identification
5. DUI (distinct user identification**)**

Proactive Mechanism: Proactive mechanisms that enforce correct mappings during the activities of each visitor. This strategies aim at differentiating the users before or during the page request [5].

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 heuristics: Reactive heuristics that perform the mappings a posteriori. Reactive strategies attempt to associate individuals with the log entries after the

log is written. 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 [5].

IP Address Based Identification: The unique user is identified by using IP address, User agent and Referred URL field**.** The users with the same IP address field are considered to be same. If IP address is same but user agent is different than each different agent represent the different user.

The strategy of User Identification based on the log entries without considering the topology structure of site.

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.

Cookie-based user identification: The extended log format of the web servers does not provide enough information, a more complex approach is needed for recording more information about the different users. This is done by using cookies. The process of recording page visits is as follows [2].

A. Cookie Handling

For tracking the users' behavior the log files are extended with cookies and some other fields as well. Cookies are the most common way of client side data storing, as the web sites that provide the pages (and write the log file) send a data packet (cookie) to the client's browser at the first visit, and then this data is sent back to the server each the user navigates to one of the pages of the same site. Downloading a unique identifier in the data packet for each client will allow recording additional information on the server side. sites. For this reason the logging is done on a central server. In this way not only the user's behavior on a single web site can be observed but on several sites that are part of the investigation. Thus, comparative measurements for the different web sites can be achieved as well. This is done by using third party cookies (denoted with C3). Technically it is realized by embedding a small picture in all the pages of all web sites that are part of the investigation. This small picture (usually a 1*1 pixel GIF in background color) is a web reference to the logging server, and it is downloaded each time a page request is coming to the content provider. This reference places a cookie on the client's browser by the first request. As this C3 belongs the the central server (domain), it will be unique for the browser, and will remain the same for all web sites that are involved in the investigation. It means that the browser uses the same cookie regardless to the investigated site. Afterwards, when the given browser downloads the same page or other page that is involved in the investigation, the cookie is sent back, and the server writes the appropriate record to the log.

The main problem with this type of information collection is that security software and also the users themselves often delete third party cookies. It distorts the results because in this case new cookie will be assigned to the client, and it will be considered as a new user by the system. To overcome this difficulty the web sites that are involved in the measurement use first party cookies (denoted with C1) as well. In this case the central server places a small script in the page, and this script generates the first party cookie in the name of the investigated web site. The security software do not delete first party cookies; because they are considered as an integral part of the site (some user specific information is stored in C1 by the site like personal settings, persistent information or baskets of an on-line store etc.). This method ensures that the user's browser contains a unique C1 for each web site, and one single C3 for all the web sites. When the client's browser resolves the reference embedded in the page requested and turns to the server it appears on the logging serves with the C3 belonging to the user along with the C1 belonging to the pair of user and the web page downloaded by the user. In this way the different third party cookies that are deleted and regenerated for the same user can be connected by the first party cookie belonging to the given user. In this way the user behavior can be tracked more precisely. For tracking the individuals behind the web users a kind of authentication is needed. Some sites requires authentication like web mail systems, web stores, forums and so on.

DUI (distinct user identification): Distinct user identification technique which enhancement of pre-processing steps of web log usage data in data mining [4]. We use two preprocessing technique combine within one preprocessing step time of user identification we find out distinct user based on their attended session time. 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. This 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. *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 improve the overall designing and performance of future access.

III. COMPARATIVE STUDY

As per literature survey we have studied many papers on the user identification through web log data which can be compare and put such a graph which will be useful for the future research. There are various paper techniques are available in this area. So we are doing comparative study of user identification through web log data based on the advantages and disadvantages.

User Identification Techniques	Advantage	Disadvantage
Proactive Mechanism	• Differentiating the users before or during the page require	• Site structure based which reduced the efficiency of identification
Reactive heuristics	• Performs mappings posteriori.	
IP Address Based Identification	• User Identification without considering topology structure of site.	• Less accurate
Cookie-based user identification	• Records more information about the different user	• Software Security • Users themselves often delete third party cookies
DUI (distinct user identification)	• Very efficient as compare to other identification techniques	

IV. CONCLUSION

Web usage and data mining to find patterns is a growing area with the growth of Web-based applications. Application of web usage data can be used to better understand web usage, and apply this specific knowledge to better serve users. User identification an important issue is how exactly the users have to be distinguished. In this paper we have discussed different user identification Techniques with their comparative study. DUI algorithm is very efficient as compare to other identification techniques.

REFERENCES

[1] 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 23.

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

[3] 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

[4] Sheetal A Raiyani et al , International Journal of Computer Science & Communication Networks,Vol 2(4), 526-530 526 ISSN: 2249-5789

[5] Berendt, B., Mobasher, B., Spiliopoulou, M., and Wiltshire, J. 2001. Mea- suring the accuracy of sessionizers for web usage analysis. Proceedings of the Workshop on Web Mining, First SIAM International Conference on Data Mining, Chicago.

[6] International Journal of Engineering and Innovative Technology (IJEIT) Volume 1, Issue 4, April 2012 ISSN: 2277-3754

[7] 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.