IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 3, Issue 4, Aug-Sept, 2015
ISSN: 2320 – 8791 (Impact Factor: 2.317)
www.ijreat.org

Automated Tool For Web User Identification

Bhawesh Kumar Thakur [1], Syed Qamar Abbas[2], Mohd. Rizwan Beg[3], Sheenu Rizvi[4]

[1] Department of Computer Science & Engineering, BIET, UPTU, Lucknow, UP,226021, India
[2]Department of Computer Science & Engineering, AIMT, UPTU, Lucknow, UP, 226001, India
[3]Department of Computer Science & Engineering, RBGI, Agra, Agra, UP, 282001, India
[4]Department of Computer Science & Engineering, Amity University, Lucknow, UP, 226001, India

Abstract

Nowadays identifying user on the web correctly is required for satisfying their needs exactly and for better user satisfaction. In our approach a web user identification can be done automatically by taking IP address of the user from web usage data i.e. Web Log file as well as managing a server based agent that is responsible for retrieving current user's Windows account information of the client. The obtained user information can be utilized to recommend new and related web pages to the web user that can be useful for making a better and efficient web personalization system. ***Keywords:*** *web usage mining, Web Log, Preprocessing, Web User, Cookie*

1. Introduction

In the era of rapidly developed Internet based applications and huge number of web users implies a wide environment for research in the area of web mining. Providing contents related to current web surfing makes the user environment more comfortable. It is one of the vital issues that brought a major attention to the researchers.

Data Mining is basically uncovering interesting data patterns hidden in large databases. Thus, data mining should have been more appropriately named "knowledge mining from data". Alternatively, it has been called exploratory data analysis, data driven discovery, and deductive learning.

Web mining is one of the areas of Data Mining that refers to the use of data mining techniques to automatically retrieve, extract and evaluate information for knowledge discovery from Web documents and services. Web data is typically unlabelled, distributed, heterogeneous, semi-structured, time varying, and high dimensional.

Web Usage Mining as name suggest gives opportunity to use log file data by which web user environment can be personalized in effective manner. Nowadays, various data mining techniques have been successfully applied to Web access logs to extract useful information. Among them, clustering allows us to group together clients or data items that have similar characteristics [4] and user profiling provides the important information describing who is the user and how they behave [13].

Web personalization is defined as any action that adapts the information or services provided by a Web site to the needs of a particular user or a set of users, taking advantage of the knowledge gained from the users' navigational behavior and individual interests, in combination with the content and the structure of the Web site. The objective of a Web personalization system is to "provide users with the information they want or need, without expecting from them to ask for it explicitly".

Principal elements of Web personalization include:-
a) categorization and preprocessing of Web data,
b) extraction of correlations between and across different kinds of such data,
c) determination of the actions that should be recommended by such a personalization system.

Web data are those that can be collected and used in the context of Web personalization. These data are classified in four categories [5]:

- Content data are presented to the end-user appropriately structured. They can be simple text, images, or structured data, such as information retrieved from databases.
- Structure data represent the way content is organized. They can be either data entities used within a Web page, such as HTML or XML tags, or data entities used to put a Web site together, such as hyperlinks connecting one page to another.
- Usage data represent a Web site's usage, such as a visitor's IP address, time and date of access, complete path

accessed, referrers' address, and other attributes that can be included in a Web access log.

• User profile data provide information about the users of a Web site. A user profile contains demographic information (e.g. name, age, country, etc.) for each user of a Web site, as well as information about users' interests and preferences. It also represents user's short-term or long term interests and is usually built as a concept hierarchy. The query that user issued and document that user browsed are categorized into concept hierarchies that are accumulated to generate a user profile [16].

Such information is acquired through registration forms or questionnaires, or can be inferred by analyzing Web usage logs. Several approaches like unified probabilistic model can be used to handle difficulties related to these data [15]. Constructing accurate and comprehensive user profiles of individual users is one of the key issues in developing personalization applications [8].

2. Web Usage Mining Tasks

Web Usage Mining of the data generated by the users interactions with the Web, typically represented as Web server access logs, user profiles, user queries and mouse-clicks. Web Usage Mining (WUM) can be used to discover interesting knowledge from web server log. The access log files of a Web server contain a lot of details about users' on-site behavior [9, 11]. This includes trend analysis (of the Web dynamics information), and Web access association/sequential pattern analysis. Also it captures interesting and uninteresting web pages from user browsing behavior [12]. The web usage mining uses various tasks by which efficient and effective services to the web users can be provided [1]:

• Data assembling: This is done by the web servers. Data can also be collected at client sides.

• Data cleaning: In web usage mining, there may be possibilities to have unwanted data recorded in the log file that is not useful for the further process, or even misleading or faulty. These records have to be corrected or removed.

• User identification: In this step the unique users are distinguished, and as a result, the different users are identified.

The objective of identification process is to extract the different users from the web log. User identification is a one of the important activity of WUM. Step of user identification filters and labels unique users of the log data [10]. The main focus of any web site is to make profit or provide user satisfaction, so user identification is required to make further strategy. This refinement is either is on structure or in contents of web site. Session identification and path completion are also useful for analysis.

Traditionally user identification activity is performed based on these rules [6, 7]:

1) Different IP address refers to different users.

2) The same IP address with different operating system or different browser should be considering as different user.

3) If the IP address, 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.

• Several methods for User Identification used [6]:

A. By IP Address:

This is frequently used technique for user identification. IP address is unique identification of computer in the Internet. Using the IP address user be can identified but this is not true in all the cases. It is assumed that the user having the same IP address is same.

B. By user registration data:

By user's registration data like, name, address, phone detail, etc, more reliably user can be identified provided all information filled by user is correct.

C. By cookies:

Cookies are the part of information which keeps on the client's computer for given amount of time. Cookies are used for fast access to web site.

The user can recognized with the following meaning-

1. A particular person.

2. A Specific category :

a. Working Profile: Student, Customer, Business Person, etc.

b. Age Group: Children, teenagers, youngsters, etc

c. Nature: Religious, Adventurous etc

d. Hobbies: Music, Sports, etc.

e. Temporal: Morning user, afternoon user, evening user, late night user etc.

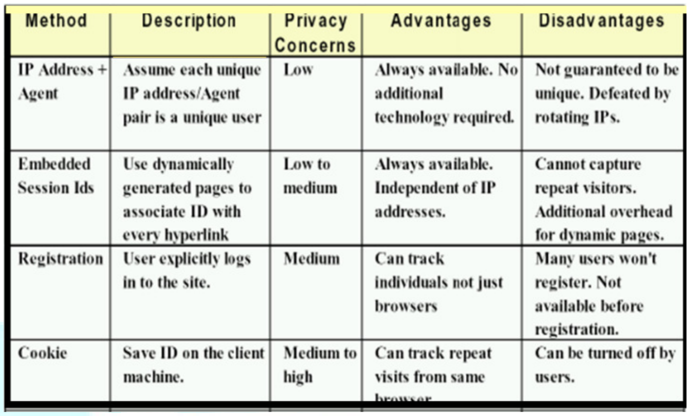
In WUM the meaning of user is mostly belonging to the specific category of user on the basic of scenario not refer to specific user.

Session identification: A session can be identified as a sequence of activities performed by a user when she/he is navigating through a given site within consecutive time period. Identification of a session is also a complex activity. Since the user may have different browsing goals each time he/she accesses the site and since sessions are easier to identify in log files than users, sessions can also be used as instances (instead of users) [14].

• Feature extraction: In this step only those fields are selected, that are relevant for further processing.
• Data transformation: In this step, the data is transformed in such a way that the data mining task can use it.
• Data mining activities execution: Data mining activities are then executed to extract the knowledge from the data
• Result interpretation and perception: The extracted result is used for taking further decisions.

3. Existing Web User Identification Approaches

In order to personalize a Web site, the system should be able to distinguish between different users or groups of users [3]. For the web user identification an important issue is how exactly the users have to be distinguished. It depends on the task for the mining process is executed. In certain cases the users are identified only with their IP addresses [2]. Another approach is based on cookie that can be transferred to user machine by which user's logged in information can be retrieved and send it to web log server for user identification. In other cases some heuristics are used for better identification of the users. Different web user identification 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. Recently, many research projects are dealing with Web usage mining and Web personalization areas. Most of the efforts focus on extracting useful patterns and rules using data mining techniques in order to understand the users' navigational behavior, so that decisions concerning site restructuring or modification can then be made by humans. In several cases, a recommendation engine helps the user navigate through a site. Some of the more advanced systems provide much more functionality, introducing the notion of adaptive Web sites and providing means of dynamically changing a site's structure.



Method	Description	Privacy Concerns	Advantages	Disadvantages
IP Address + Agent	Assume each unique IP address/Agent pair is a unique user	Low	Always available. No additional technology required.	Not guaranteed to be unique. Defeated by rotating IPs.
Embedded Session Ids	Use dynamically generated pages to associate ID with every hyperlink	Low to medium	Always available. Independent of IP addresses.	Cannot capture repeat visitors. Additional overhead for dynamic pages.
Registration	User explicitly logs in to the site.	Medium	Can track individuals not just browsers	Many users won't register. Not available before registration.
Cookie	Save ID on the client machine.	Medium to high	Can track repeat visits from same browser	Can be turned off by users.

Table 1: Comparison of different user identification

4. Implementation of Proposed Web User Identification System

• log server sends a cookie to content provider for the purpose of user identification.
• Client request page view to content provider very first time.
• Content provider sent a requested page with a log server cookie to the client.
• Cookies at Client collect Windows's login information i.e. user name, IP Address and login time from client system and send it to the server.
• Client request the content provider for the pages.
• Content provider sent requested web page info to the log server.
• After that log server provide recommendations to the specific user.

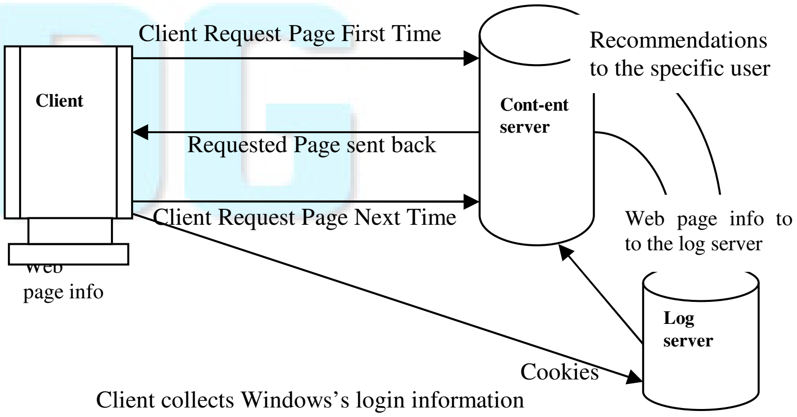

Client Request Page First Time
Recommendations to the specific user
Client
Cont-ent server
Requested Page sent back
Client Request Page Next Time
Web page info to to the log server
Web page info
Log server
Cookies
Client collects Windows's login information

Fig. 1: Automatic Web User Identification System

4.1 Algorithm for user Identification:

Step-1 (Definition 1)
A Log is a set of triplets {(u1,p1,t1)…(un,pn,tn)} where ui € U (set of users), pi €P (set of pages/clicks), and ti is a time stamp.
Step-2 Data cleaning
Elimination of the items (URLs) deemed irrelevant in data set. (By checking the suffix of the URL name).
Step-3 User Identification
- Log server send cookie to content provider.
- Client request for required page.
- Content provider send log server's cookie along with requested page.
- Cookies at Client collect Windows's login information i.e. user name, IP Address and login time from client system and send it to the content provider.
- IP addresses with Windows's login information from web log server are used to identify unique web users.

Step-4 Formatting
Apply final preparation module to format properly the session file.
Step-5 Web Personalization techniques are used to provide recommendation sets for the identified users.
Step-6 Outputs of the system are used by the user conveniently and in an user-friendly environment.

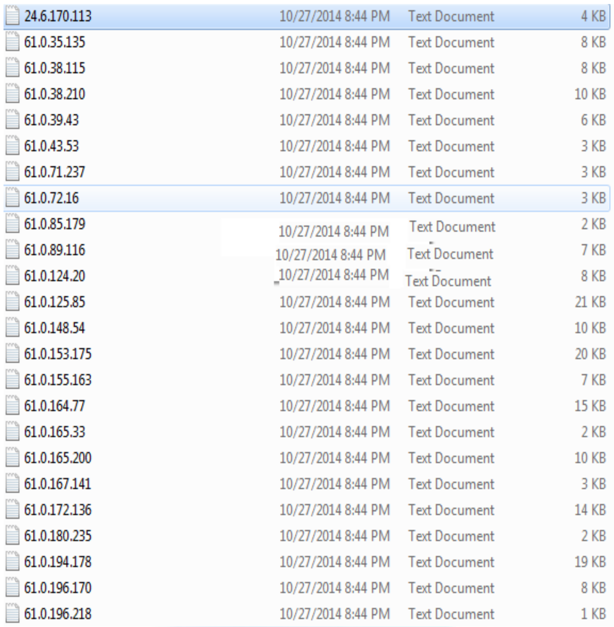
4.2 Psuedocode for User Identification

```
public class LogSplit {
          public static void main(String[] args)
{
                    FileInputStream fin=null;
                    FileOutputStream fw=null;
                              try {
                         fin=new
FileInputStream("Log2.txt");
                         //BufferedReader
br=new                     BufferedReader(new
InputStreamReader(fin));
                         String line="";
                         String oldName="";
```

```
while((line=readLine(fin))!=null)
                              {
          if(line.startsWith("#"))continue;
                                        String
splits[]=line.split("\\s");
                                        String
fileName=splits[2];
          if(!oldName.equals(fileName))
                                        {
oldName=fileName;
          if(fw!=null)fw.close();
          fw=new
FileOutputStream("output/"+fileName+".txt
",true);
                                        }
          fw.write(line.getBytes());
          fw.write("\n".getBytes());
          //if(i++==4)break;
                              }
                    } catch (Exception e) {
                         //      TODO      Auto-
generated catch block
                              e.printStackTrace();
                    }
}
          private          static          String
readLine(InputStream               is)throws
IOException
          {
                    StringBuffer          line=new
StringBuffer();
                    while(is.available()>0)
                    {
                    char ch=(char)is.read();
                    if(ch=='\n')
                    {
                              return line.toString();
                    }
                    line.append(ch);
                    }
                    return null;
}}
```

4.3 Identified Users

From MCU web server log file following users are identified who interact with web server in one day i.e. 27-10-2014.



24.6.170.113	10/27/2014 8:44 PM	Text Document	4 KB
61.0.35.135	10/27/2014 8:44 PM	Text Document	8 KB
61.0.38.115	10/27/2014 8:44 PM	Text Document	8 KB
61.0.38.210	10/27/2014 8:44 PM	Text Document	10 KB
61.0.39.43	10/27/2014 8:44 PM	Text Document	6 KB
61.0.43.53	10/27/2014 8:44 PM	Text Document	3 KB
61.0.71.237	10/27/2014 8:44 PM	Text Document	3 KB
61.0.72.16	10/27/2014 8:44 PM	Text Document	3 KB
61.0.85.179	10/27/2014 8:44 PM	Text Document	2 KB
61.0.89.116	10/27/2014 8:44 PM	Text Document	7 KB
61.0.124.20	10/27/2014 8:44 PM	Text Document	8 KB
61.0.125.85	10/27/2014 8:44 PM	Text Document	21 KB
61.0.148.54	10/27/2014 8:44 PM	Text Document	10 KB
61.0.153.175	10/27/2014 8:44 PM	Text Document	20 KB
61.0.155.163	10/27/2014 8:44 PM	Text Document	7 KB
61.0.164.77	10/27/2014 8:44 PM	Text Document	15 KB
61.0.165.33	10/27/2014 8:44 PM	Text Document	2 KB
61.0.165.200	10/27/2014 8:44 PM	Text Document	10 KB
61.0.167.141	10/27/2014 8:44 PM	Text Document	3 KB
61.0.172.136	10/27/2014 8:44 PM	Text Document	14 KB
61.0.180.235	10/27/2014 8:44 PM	Text Document	2 KB
61.0.194.178	10/27/2014 8:44 PM	Text Document	19 KB
61.0.196.170	10/27/2014 8:44 PM	Text Document	8 KB
61.0.196.218	10/27/2014 8:44 PM	Text Document	1 KB

Fig. 2 Identified Users

4.4 Transaction For Identified User- 24.6.170.113

```
2014-10-27 19:27:00 24.6.170.113 - W3SVC195 NS
69.41.233.13 80 GET /AboutUniversity.htm - 200 0
10668 144 211 HTTP/1.1 www.mcrpv.ac.in
TerrawizBot/1.0+(+http://www.terrawiz.com/bot.html) -
-
2014-10-27 20:06:45 24.6.170.113 - W3SVC195 NS
69.41.233.13 80 GET /index.htm - 200 0 21738 125 281
HTTP/1.1 www.mcrpv.ac.in
TerrawizBot/1.0+(+http://www.terrawiz.com/bot.html) -
-
2014-10-27 20:14:14 24.6.170.113 - W3SVC195 NS
69.41.233.13 80 GET
/Important_Notice_Syllabus+for+the+BCA+course+for
+the+session+2004-2007.htm - 200 0 3392 215 181
HTTP/1.1 www.mcrpv.ac.in
TerrawizBot/1.0+(+http://www.terrawiz.com/bot.html) -
```

```
2014-10-27 20:31:08 24.6.170.113 - W3SVC195
NS 69.41.233.13 80 GET /New_and_Updates.htm
- 200 0 11622 144 220 HTTP/1.1
www.mcrpv.ac.in
TerrawizBot/1.0+(+http://www.terrawiz.com/bot.ht
ml) - -
```

Fig. 3 Snapshot of transaction for identified user- 24.6.170.113

5. Conclusions

Detecting exact web user for the purpose of providing web content in personalization is one of the trivial issues that can be handled using various approaches. Identifying correct user on the web makes sense for managing user needs in proper manner as well as for better user satisfaction. In our approach a web user identification system is proposed that can automatically identify user by capturing IP address of the user from web usage data i.e. Web Log file as well as managing a server based agent that is responsible for retrieving current user's Windows account information from the client. The identified user information can be utilized to recommend new and related web pages to the web user that can be useful for making a personalized web interaction of user.

References

[1] Pabarskaite, Z. and Raudys, A., (2007), 'A process of knowledge discovery from web log data: Systematization and critical review', Journal of Intelligent Informatin Systems, Vol. 28. No. 1. , pp 79-104.

[2] Gery, M. and Haddad, H., (2003), 'Evaluation of Web Usage Mining Approaches for User's Next Request Prediction', Fifth International Workshop on Web Information and Data Management (WIDM'03), pp 74-81.

[3] Eirinaki, Magdalini And Vazirgiannis, Michalis, (2003), 'Web Mining for Web Personalization', ACM Transactions on Internet Technology, Vol. 3, No. 1, pp 1–27.

[4] HuaXu, Lin and Liu, Hong, (2010) , 'Web User Clustering Analysis based on KMeans Algorithm', IEEE International Conference on Information, Networking and Automation (ICINA), pp 6-9.

[5] Srivastava, Jaideep and Cooley, Robert and Deshpande, Mukund and Tan, Pang-Ning, (2000), 'Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data', ACM SIGKDD, Vol.1(2), pp 12-23.

[6] Bakariya, Brijesh and Mohbey, Krishna K. and Thakur, G.S., (2011), 'An Inclusive Survey on Data Preprocessing Methods Used in Web Usage Mining', Springer.

[7] Singh, Satpal and Badhe, Vivek, (2014), 'An Exclusive Survey on Web Usage Mining For User Identification', International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 11, pp 6582-6589.

[8] Yang, Yinghui (Catherine), (2010), 'Web user behavioral profiling for user identification', Elsevier Decision Support Systems 49, pp 261-271.

[9] Patel, Priyanka and Parmar, Mitixa, (2014), 'A Review on User Session Identification through Web Server Log', International Journal of Computer Science and Information Technologies, Vol. 5 (1), pp 146-148.

[10] Iváncsy, Renáta and Juhász, Sándor, (2007), 'Analysis of Web User Identification Methods', International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:1, pp 2995-3002.

[11] Chan, Philip K., (2002), 'Constructing Web User Profiles: A Non-invasive Learning Approach', Web Usage Analysis and User Profiling, pp 39-55.

[12] Hawalah, A and Fasli, M., (2011), 'A hybrid re-ranking algorithm based on ontological user profiles', IEEE Computer Science and Electronic Engineering Conference (CEEC), pp 50-55.

[13] Adomavicius, Gediminas and Tuzhilin, Alexander, (1999), 'User Profiling in Personalization Applications through Rule Discovery and Validation', KDD '99 Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pp 377-381.

[14] Grčar, Miha, (2004), 'User Profiling: Web Usage Mining', 7th International Multiconference Information Society IS.

[15] Tang, Jie, and Yao, Limin and Zhang, Duo and Zhang Jing (2010), 'A Combination Approach to Web User Profiling', ACM Transactions on Knowledge Discovery from Data (TKDD), Volume 5 Issue 1.

[16] Yu, Jie and Liu, Fangfang, (2010), 'Mining user context based on interactive computing for personalized Web search', 2nd International Conference on Computer Engineering and Technology (ICCET) (Volume:2), pp 209-214.