ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

An Exclusive Survey on Web Usage Mining For User Identification

Satpal Singh[1], Vivek Badhe[2]

M.Tech. Scholar, Dept. of CSE, Gyan Ganga College of Technology, Jabalpur, MP, India[1]

Assistant Professor, Dept. of CSE, Gyan Ganga College of Technology, Jabalpur, MP, India[2]

ABSTRACT: Web mining has been explored to a vast degree with different techniques that has been proposed for a variety of applications. Most research on web mining has been done on "data-centric" point of view and a few works has been done on "user-centric" view. In this paper we explored the definition of web-user and also examined the various dimensions of temporal web mining.

In web usage mining, web log file plays an important roll, some fields of the web log file are used frequently and some other fields like *DateTime* are rarely used. They are in advertently removed in the cleaning session. This paper emphasizes the importance of these fields, which can be very useful for user identification process.

We study in particular the behaviour of web-usage data over a period of time. In this paper we have suggested the different view point for finding the web-user on the basic of temporal approach. Such kind of analysis could be useful for target marketing based on time or for web services optimization.

KEYWORDS: Web mining, Web usage mining, Web log, User identification, Temporal web mining

I. INTRODUCTION

Data mining is the process of analysing data from different perspectives and summarizing it into useful information. Web mining is the application of data mining techniques to extract knowledge from Web data. Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined. They are web content mining, web usage mining and web structure mining.

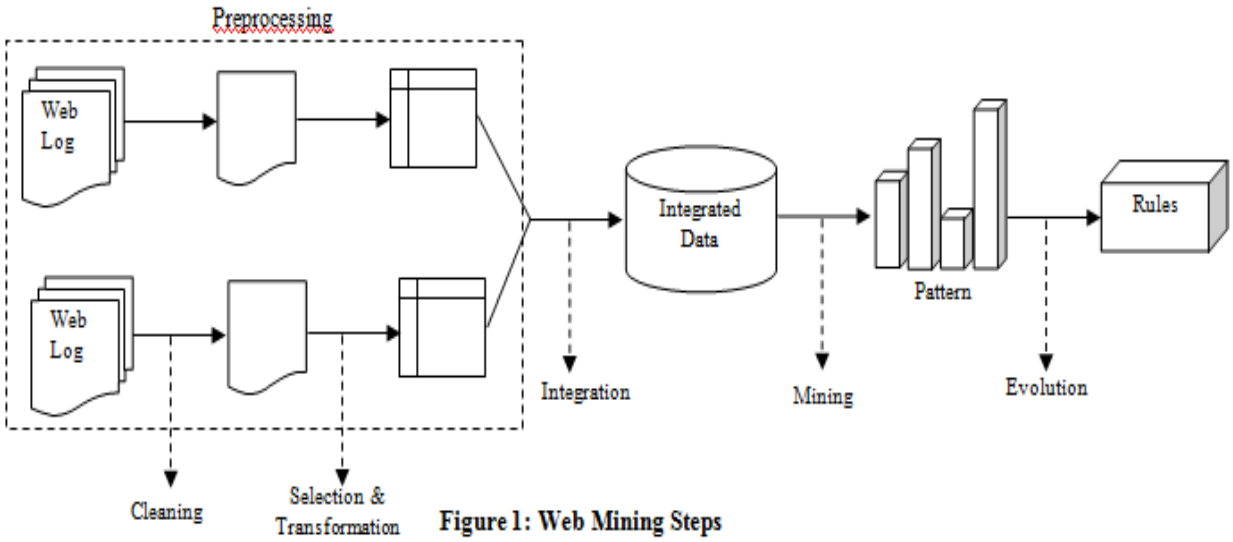

Figure 1: Web Mining Steps

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

Data Mining: Extract the patterns from the large amount of data is called Data Mining. Data Mining [11] is the most important step of KDD process. There are different types of mining are used, i.e. Web mining, Sequence mining, Temporal Mining, mining, Multimedia and Spatial mining. There are three fundamental methods are available for mining i.e. Association Rule Mining, Classification and Clustering.

Web Mining: The mining apply on the data witch available on the Web is called Web Mining. Web Mining techniques to make the web more useful and more profitable (for some) and to increase the efficiency of our interaction with the web. Web Mining broadly divided into three distinct categories according to the kinds of data to be mined. Figure 1.

a. *Web Content Mining:* Web content mining is the process of extracting useful information from the contents of web documents. Web Content Mining deals with the discovery of useful information from the web contents or data or documents or services.
b. *Web Structure Mining:* The structure of a typical web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Web structure mining is the process of discovering structure information from the web.
c. *Web Usage Mining:* Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications. Usage data captures the identity or origin of web users along with their browsing behaviour at a web site. There are four stages in web usage mining.

Web Data: One of the most important steps in knowledge discovery in databases is to construct a proper target data set for the data mining task. In Web data mining, data can be gathered from Web servers, client sites, and proxy server or obtained from organization's database. Different type of data is collected from different location. There are many types of data that can be used in Web Mining [1].

a. *Web Content*
The data that is present on the Web pages which provide information to the users. Some examples of Web Content data are text, HTML, audio, video, images, etc.
b. *Web Structure*
The Web pages are connected with each other through hyperlinks i.e. various HTML tags used to link one page to another and one Web site to another Web site.
c. *Web Usage*
These data reflect the usage of Web and are collected on Web servers, proxy server, and client browser with IP address, date, time etc. This type of data is auto generated by the web server and well known as Web-Log and the file which contains that data is called web-log file and it semi-structured text file. *Refer Table 1 and 2.*
d. *Web User Profile*
The data that provides demographic information about users of the Web sites, i.e. user registration data and customers profile information.

Data Collection: Users log data is collected from various sources like server side, client side, and proxy servers and so on [1]
Data Collection is the first step in web usage mining process. It consists of gathering the relevant web data. Data source can be collected at the server-side, client-side, proxy servers, or obtain from an organization's database, which contains business data or consolidated Web data.

Server level collection collects client requests and stored in the server as web logs. Web server logs are plain text that is independent from server platform. Most of the web servers follow common log format as "IP Address, username/password date/timestamp, URL, version, status-code, bytes-sent" Some servers follow

Client Level Collection is advantageous than server side since it overcomes both the caching and session identification problems. Browsers are modified to record the browsing behaviours.

Proxy level collection is the data collected from intermediate server between browsers and web servers. Proxy traces may reveal the actual HTTP requests from multiple clients to multiple Web servers. This may serve as a data source for characterizing the browsing behaviour of a group of anonymous users sharing a common proxy server.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

5.26.149.185 - - [04/Nov/2002:01:01:53 +0000]"GET /ivsat.htm HTTP/1.1" 200 9430 "http://www.yahoo.com/index.htm" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)"

65.26.149.185 - - [04/Nov/2002:02:15:03 +0000]"GET /901-342s.jpg HTTP/1.1" 200 8600 "http://www.satsig.net/ivsat.htm" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)" 65.26.149.185 - - [04/Nov/2002:01:51:54 +0000] "GET /pas1rkuh.gif HTTP/1.1" 200 4189 "http://www.satsig.net/ivsat.htm" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)"

65.26.149.185 - - [04/Nov/2002:04:45:35 +0000] "GET /nss7kwas.jpg HTTP/1.1" 200 6271 "http://www.satsig.net/ivsat.htm" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)" 65.26.149.185 - - [04/Nov/2002:01:51:54 +0000] "GET /asiak2.gif HTTP/1.1" 200 6560 "http://www.satsig.net/ivsat.htm" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)"

65.26.149.185 - - [04/Nov/2002:01:51:54 +0000] "GET /ab2_eu3.gif HTTP/1.1" 200 6635 "http://www.satsig.net/ivsat.htm" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)" 66.32.2.122 - - [04/Nov/2002:01:52:55 +0000] "GET /ssazelm.htm HTTP/1.1" 304 - "http://www.google.com/search?hl=en&lr=&ie=UTF-8&oe=UTF-8&as_qdr=all&q=satellite+signal+meter+aim" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0; Q312461)"

66.32.2.122 - - [04/Nov/2002:10:48:21 +0000]"GET /sf-95-3.gif HTTP/1.1" 304 - "http://www.satsig.net/ssazelm.htm" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0; Q312461)" 24.43.169.115 - - [04/Nov/2002:01:53:02 +0000] "GET /ssazelm.htm HTTP/1.1" 200 11623 "http://www.google.ca/search?q=%22Free+to+Air%22%2Bsatellite+dish&hl=en&lr=&ie=UTF-8&oe=UTF-8&start=10&sa=N" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.0.3705)"

24.43.169.115 - - [04/Nov/2002:11:51:45 +0000]"GET /sf-95-3.gif HTTP/1.1" 200 3536 "http://www.satsig.net/ssazelm.htm" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.0.3705)" 64.130.130.17 - - [04/Nov/2002:01:55:13 +0000] "GET /ssazelm.htm HTTP/1.0" 200 11857 "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0)" 64.130.130.17 - - [04/Nov/2002:01:55:14 +0000] "GET /sf-95-3.gif HTTP/1.0" 200 3536 "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0)"

Client IP	Access Date And Time	Method	URL	Protocol	Status	Bytes	User Agent
5.26.149.185	[04/Nov/2002:01:01:53 +0000]	GET	http://www.yahoo.com/index.htm	HTTP/1.1	200	9430	Mozila4.0
65.26.149.185	[04/Nov/2002:02:15:03 +0000]	GET	satsig.net/ivsat.htm	HTTP/1.1	200	8600	Mozila4.0
65.26.149.185	[04/Nov/2002:04:45:35 +0000]	GET	www.satsig.net/ivsat.htm	HTTP/1.1	200	6271	Mozila4.0
65.26.149.185	[04/Nov/2002:06:01:33 +0000]	GET	www.satsig.net/ivsat.htm	HTTP/1.1	200	6635	Mozila4.0
66.32.2.122	[04/Nov/2002:10:48:21 +0000]	GET	satsig.net/ssazelm.htm	HTTP/1.1	304	0	Mozila4.0
24.43.169.115	[04/Nov/2002:11:51:45 +0000]	GET	www.satsig.net/ssazelm.htm	HTTP/1.1	200	3536	Mozila4.0

IP Address: IP address of the remote host.
Date: Date and time of the request.
Request/ Method: The request line exactly as it came from the client.
URL: The URL the client was on before requesting your URL.
Protocol: The protocol user web page is running; HTTP / HTTPS
Status: The HTTP response code returned to the client.
Bytes: The number of bytes transferred.
User Agent: The software the client claims to be using.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

Cookies are unique ID generated by the web server for individual client browsers and it automatically tracks the site visitors. When the user visits next time the request is send back to the web server along with ID. However if the user wishes for privacy and security, they can disable the browser option for accepting cookies.

Explicit User Input data is collected through registration forms and provides important personal and demographic information and preferences. However, this data is not reliable since there are chances of incorrect data or users neglect those sites.
The information available in the web is heterogeneous and unstructured. Therefore, the preprocessing phase is a prerequisite for discovering patterns. The goal of preprocessing is to transform the raw click stream data into a set of user profiles.

II. RELATED WORK

In WUM research work is continually progress in preprocessing and user identification. As we know preprocessing is very important task of web-mining. The results of mining are depend upon the preprocessing and also directed to the mining process. Researcher are introduced the various methods and algorithms for preprocessing. Similarly, user identification is also a very difficult task, researcher are taking this problem in deferent way and also provide the different solution of above but research of exact solution is still going on.

- Jaideep Shrivastava et. al. [1] publishes a very popular and important paper which includes in most of the papers as a reference, is a part of most of the papers also discuss the problem of user identification.

- Reddy et. al. [2] proposed the model for data preprocessing, as per paper this model works for data cleaning, unique users and session record, but still there is a problem in quality of data, accuracy metric of the user identification and the session identification and applying the results of the preprocessing to discover patterns.

- Chintan R. Varnagar et. al. [4] wrote, most of the systems, architecture that was implemented or proposed considers either client side or server side log data. In future a system could be build that considers and exploit the usefulness of both client side and server side log data, to produce result that are more efficient and batter match with empirical observations.

- Brijesh Bakaria et. al. [3], publishes is a survey paper which of 2013, discuss till date there is no concrete solution is available for user identification.

- Liu Kewen[5], proposed the algorithm for data cleaning but discuss the problem of user identification. But it is difficult to take a challenge of over TB level data.

- Sheetal A. Raiyani et. al. [6] , proposed the algorithm called DUI (Distinct User Identification) as per author 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

- V. Sujatha et. al.[7], proposed the algorithm based on Pattern using Clustering & Classification (PUCC), This step of PUCC focuses on separating the potential users from others. Suneetha and Krishnamoorthy (2010) used decision tree classification using C4.5 algorithm to identify interested users. They use a set of decision rules for this purpose. The algorithm worked efficiently in identifying potential users, but had the drawback that it completely ignored the entries made by network robots. Search engines normally use network robots to crawl through the web pages to collect information. The number of records created by these robots in a log file is extremely high and has a negative impact while discovering navigation pattern. This problem is solved in this

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

paper by identifying the robot entries first before segmenting the user groups into potential and not-potential users.

- Hongzhou Sha et. al. [8] proposed method EPLogCleaner: Improving Data Quality of Enterprise Proxy Logs for Efficient Web Usage Mining, Experiment results show that EPLogCleaner can filter out more than 30% URL requests which cannot be filtered by traditional data cleaning methods for proxy logs. But not all filtered data is valuable and relevant. Some keep alive links add timestamp into their URLs, so their prefix cannot be added directly to our prefix library simply by the threshold. It made some irrelevant and useless data stay in the final result. Besides, the design of threshold and the estimation method of precision rate are relatively simple. Next, we will first analysis the timestamp information in the URL, capture its characteristics in order to obtain higher filtering rate. Moreover, we will improve the design of threshold and the estimation method of precision rate in order to make the experimental results much more accurate and reliable.
- Mofreh Hogo et. el. [9] introduces the temporal web usage mining of web users on educational web site, using the adapted Kohonen SOM based on rough set properties
- Sourabh Jain et. at. [10] presented paper is a review in temporal data mining and the fuzzy association rule in order to get the required data fastly and efficiently as well.

III. PREPROCESSING

The data preprocessing is the initial step in the data mining process. The above mention data sources are available but web log file is primary data source of web data mining. Web data mining include data cleaning, user identification, session identification, path completion [2]. Refer Figure 2.

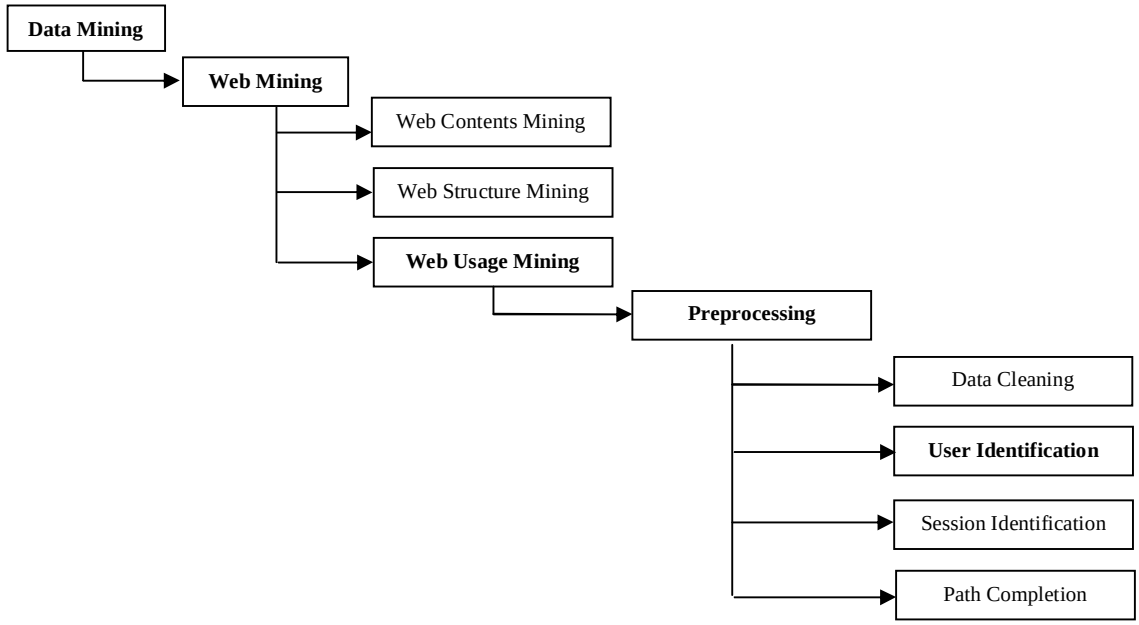

Figure 2: Object Hierarchy

A. Data Cleaning

Data Cleaning is a process of removing noise, unused and irrelevant items such as jpeg, gif files or sound files and references due to spider navigations. Improved data quality improves the analysis on it. The HTTP protocol requires a separate connection for every request from the web server. If a user request to view a particular page along with server log entries graphics and scripts are download in addition to the HTML file. An exception case is Art gallery site where images are more important. When a user download a particular page then there are different elements are also downloaded with pages like graphics and scripts. In server log entries these all element details are stored. In most cases,

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

only the log entry of the HTML file request is relevant and should be kept for the user session file then the Solution for that problem is to Eliminate some items deemed irrelevant can be reasonably accomplished by checking the suffix of URL name. All log entries with file name suffixes such as gif, jpeg etc. so that the list can be changed according to the site being analyzed[4]

B. User Identification

Identification of individual users who access a web site is an important step in web usage mining. Various methods are to be followed for identification of users. The simplest method is to assign different user id to different IP address. But in Proxy servers many users are sharing the same address and same user uses many browsers. An Extended Log Format overcomes this problem by referrer information, and a user agent. If the IP address of a user is same as previous entry and user agent is different than the user is assumed as a new user.

C. *Session Identification*

A user session can be defined as a set of pages visited by the same user within the duration of one particular visit to a web-site. A user may have a single or multiple sessions during a period. Once a user was identified, the click stream of each user is portioned into logical clusters. The method of portioning into sessions is called as Sessionization or Session Reconstruction. A transaction is defined as a subset of user session having homogenous pages. There are three methods in session reconstruction. Two methods depend on time and one on navigation in web topology.

D. Path Completion

There are chances of missing pages after constructing transactions due to proxy servers and caching problems. So missing pages are added as follows: The page request is checked whether it is directly linked to the last page or not. If there is no link with last page check the recent history. If the log record is available in recent history then it is clear that "back" button is used for caching until the page has been reached. If the referrer log is not clear, the site topology can be used for the same effect. If many pages are linked to the requested page, the closest page is the source of new request and so that page is added to the session. There are three approaches in this regard.

IV. USER IDENTIFICATION

The purpose Identification process is to find out the different users from the web. User identification is a very important task of WUM, user act as a consumer in web. Because the ultimate target of any web site is profit or user satisfaction, therefore before consumer identification we cannot make policy or strategy for betterment of our site. This betterment is either is on structure wise or in contents wise. Session identification and path completion are also very useful for analytical purpose but if this analysis process include the user identification with its session and path completion then we can get more specify and accurate results.

Methods of User Identification [3]

A. Using IP Address

This is very common heuristic technique for user identification. IP address is unique address of our computer in the Internet. Using the IP address we can identify the user but actually we not identify we are assuming that the user having the same IP address is same.

B. Using user registered data

User registered data like, user name, address, contact no, etc, comparatively more reliable source for user identification. If we considered all information filled by user is correct.

C. Using cookies

Cookies are the piece of information which stores the client's computer for specific amount of time. Cookies are basically made for fast access to web site. That means cookies can stores user's information; so using cookies we can extract the information of user.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

V. ISSUES ASSOCIATED WITH USER IDENTIFICATION

First we need to revise the definition of *user*; this word can refer the following meaning-

1. Specific person.
2. Specific category :
   a. **Working status:** Students, Customers, Business Person, Housewives etc.
   b. **Age Group :** Children, teenagers, youngsters, etc
   c. **Nature :** Religious, Adventurous etc
   d. **Hobbies:** Music, Sports, Quiz 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. The relation between different categories of users is mentioned in figure3. Specific user identification is not possible in web even in net-banking, Any person and their spouse both can share the bank account no., password and mobile no. also for OTP and share the same bank account, even sometimes in single session, but no mechanism can find the presence of second user. In bank web-log only one single user entry is recorded.

That is the reason in any online exam user identification is done by physical verification also.Second important issue is field selection from web-log. Most of the method have work on some specific fields bur some of the fields are ignored. One of the fields that is ignored is most of the use identification process is *DateTime* stamp. This field can also used for user identification. Because the group of users which use any specific site in early morning and definitely differ from the users which use that site at late night or after noon.

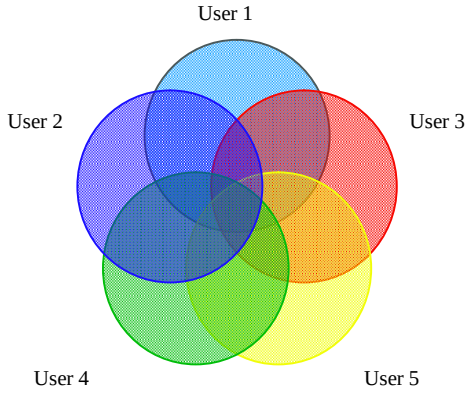

Figure 3: Relation b/w User's & *DateTime* Dimensions

DateTime is not only related to specific date or time, it has various dimensions like

1. **Specific time in a day :** 6:45:35 PM
2. **Time Period:** Office time, 10:00 AM to 5:00 PM, Examination time is 3 Hrs.
3. **Before/After:** Forenoon / Afternoon.
4. **Time slot in a day**: Early morning, evening, etc.
5. **Specific day :** 23/10/14 (Date of any occasion)
6. **Duration** : Summer Sale from 10-Jun to 20-Jun
7. **Season:** Winter, Summer, Spring, etc.
8. **Occasion:** Diwali, Christmas, New Year, etc.
9. Various combination of above; i.e. Evening of Winter season, 11:00 PM in summer. Therefore we can refer the figure3.

Therefore primarily using the *DateTime* field after mining we will get the users on the basis of date and time. These results can be useful for dynamic structure of our web site. Specially for commercial site. We observe that some of the above have crisp boundaries but some have not.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

VI. CONCLUSION AND FUTURE WORK

Hence, the paper reflects the problem of user identification in WUM and some of the papers give some methods or algorithm for this. The literature shows that some methods are specific while others have limitations. For WUM the server web log files as a dataset. Some of the fields are primarily used, like IP Address and some fields are ignored.

Web-Sessions plays the important role of in user identification. The basic assumption behind this concept is that, every session is dedicated for single user. Some researcher introduced algorithms for user identification but they do not claim the guarantee. Some other researcher gives the temporal aspect of user identification, but they work on specific area and use the vague definition of temporal and so most of the temporal dimensions are missing in WUM.

Finally we can say that various categories of users and temporal dimension will provide an avenue to the various research fields. The multiple combinations are also possible between the users with temporal dimension which can be very useful for commercial web sites.

REFERENCES

1. Jaideep Srivastava, Robert Cooleyz, Mukund Deshpande & Pang-Ning Tan, "Web Usage Mining Discovery and Applications of Usage Patterns from Web Data", ACM-SIGKDD, Jan-2000.
2. K. Sudheer Reddy, M. Kantha Reddy & V. Sitaramulu, "An Effective Data preprocessing Method for Web Usage Mining", Feb-2013, IEEE
3. Brijesh Bakariya, Krishna K. Mohbey and G.S. Thakur, "An Inclusive Survey on Data Preprocessing Methods Used in Web Usage Mining", Springer-2011.
4. Chintan R. Varnagar, Nirali N. Madhak, Trupti M. Kodinariya & Jayesh N. Rathod, "Web Usgae Mining : A Review on Process, Methods and Techniques", Feb-2013, IEEE
5. Liu Kewen, "Analysis of Preprocessing Methods for Web Usage Data", IEEE 2012
6. Sheetal A. Raiyani, Shailendra Jain and Ashwin G. Raiyani, "Advanced Preprocessing using Distinct User Identification in web log usage data", ISSN : 2278 – 1021, IJARCCE, Vol. 1, Issue 6, August 2012
7. V. Sujatha and Punithavalli, "Improved User Navigation Pattern Prediction Technique From Web Log Data", ELSEVIER-2012
8. Hongzhou Sha, Tingwen Liub, Peng Qin, Yong Sun and Qingyun Liu, "EPLogCleaner: Improving Data Quality of Enterprise Proxy Logs for Efficient Web Usage Mining", ELSEVIER-2013
9. Mofreh Hogo, Miroslav Snorek & Pawan Lingras, "Temporal Web Usage Mining", IEEE 2003.
10. Sourabh Jain, Susheel Jain & Anurag Jain, "An Assessment of Fuzzy Temporal Association Rule Mining", IJAIEM-2013
11. Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", Second Edition, ELSEVIER Inc.

BIOGRAPHY

Satpal Singh is a research scholar of M. Tech in the Gyan Ganga College of Technology Jabalpur (MP). He has completed his B.Sc. degree from Govt. Autonomous Science College, RDVV (Jabalpur), MP, after that he received Masters of Computer Application (MCA) from IGNOU. He is also done C-DAC from Kolkata and qualifies DOEACC (under AICTE) 'A' level. He is the IBM Certified Data Base Associate DB2. His area of Interest includes Data Structure, Data Mining, Soft Computing and Web Development.

Vivek Badhe is an Assistant Professor in the Department Of Computer Science and Engineering, Gyan Ganga College of Technology, Jabalpur. He received his M.Tech. in Computer Technology and Application in 2006 from Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal. He has published many research papers in various national and international journals. His areas of interest are Data Mining, Database Management System and Soft Computing.