ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 6, June 2014

Performance Enhancement of Web Server log for Distinct User Identification through different factors

Sheetal A. Raiyani[1], Rakesh Pandey[2], Shivkumar Singh Tomar[3]

M.tech, CSE, TIT, Bhopal, India [1]
Asst. Prof., CSE, TIT, Bhopal, India [2]
Asst. Prof., CSE, TIT, Bhopal, India [3]

Abstract: Nowadays millions of users daily interact with web sites around the world. Huge amount of data are being generated and these information could be very prized to the company in the field of understanding Customer's behaviours. Web usage mining is relative independent, but not isolated category, which mainly describes the techniques that discover the user's usage pattern and try to predict the user's behaviours. Web usage mining is field of data mining which deals with the originality and study of usage patterns with use of Web log data. Specifically web logs in direction to advance web based applications. Web usage mining involves of three stages, Preprocessing, Pattern discovery and Pattern analysis. Work represent a Preprocessing methodology having Data Cleaning, User Identification which is key issue in preprocessing technique phase is to identify the unique web users. Traditional User Identification is founded on the site structure and user IPaddress by using some observed rules and based on the site topology which reduced the efficiency of result. In proposed work this problem is resolve through Distinct User Identification (DUI) Technique which is based on IP address, Agent, Referred pages, status on requested page and complete session time. It can be used in unique identification of users as well as through discovery of regular access process get better designing for site modification and business intelligence. Proposed work experiments get precious accurate result which improved the overall quality of preprocessing results which more beneficial for web usage mining techniques. Cause of using different factors get well-organized preprocessed data with more time consumption.

Keywords: Web usage mining, Preprocessing, User identification, Session time, Server log

I. INTRODUCTION

One of the applications regions of data mining is *World Wide Web* (WWW) where serves as an enormous, broadly spread, global information service centre for every kind of information such as kinds of news, no. of advertisements, consumer related information, financial controlling, instruction management, Government management, e-commerce, fitness services and many other information services. Web mining has been reputable as a significant area of research. Each and every time, the web users visit the Particular Web sites they leave rich information in Web Server log. Which is fundamentally complex, heterogeneous, high dimensional and incremental in nature. Examining such data can support to determine the browsing awareness of web user. To organize this, web usage mining centres on examining the possible information from browsing designs of the users and to find the connection among the pages on study.

The main objective of web usage mining is to Capture, Ideal and Study the web log facts in such a way that it automatically determines the usage performance of web user. Subsequent the successful submission of the data mining methods in the traditional database fields, people have initiated to study the Web based data mining knowledge. The Web log mining is a technology which relates the data mining technologies to the web server log files in order to discovery the browsing patterns of users and analyse the website usage by the access of users and it can be also used to assist the site supervisors to enhance the sites structure. The Web server logs record the user's information of related to way of direction to accessing the particular web site. The usual Web server logs contain the following facts: IP address , request time , method (GET,POST,HEAD) , URL of the requested files , HTTP version , return codes , the number of bytes transferred ,the Referrer's URL and agents. Though, the data in Web logs isn't accurate because of the actuality of local cache, proxy servers and firewalls. That's why it is difficult to mark a mining directly on it and work may get particular erroneous outcomes. An quantity of investigates focused on modified service to achieve the precise technology, such as the suggested technology, data retrieval, user grouping technology, but user showing techniques are rarely declared. But, with the development and in distance study of modified service, researchers regularly realize that the excellence of modified service not only depends on the specific reference technology, search technology, but also trusts on user's favourites and other features of interest, account of its computable, while the latter is mainly important.

Resource Discovery the task of recovering the future information from Web.

- Facts Extraction: by design choosing and preprocessing precise information from the saved Web resources.

- Generalization: repeatedly determines overall patters at the both separate Web sites and across several sites.
- Investigation: analysing the extracted pattern.

Web mining is a technique to learn and examine the useful information from the Web data. The authors by [10] titles the Web includes three types of data. Data on the Web content, Web log data usage and Web structure data. R.M.Suresh and Padmajavalli, [14] considered the Web mining into specific categories Web usage mining, Web text mining and user forming mining. While today the greatest recognized categories of the Web data mining are Web content mining, Web structure mining, and Web usage mining by [2, 8, and 10]. So, it is clear that the taxonomy is based on what type of Web data to mine

II. WEB USAGE MINING

Web usage mining tries to find out useful content from the server access log at time of user interact with web site. As well as discover meaningful information at the period when interactions of the users while surfing on the Web. It concentrate on technique through work might predict user interaction. R.M.Suresh and Padmajavalli [14] work abstract strategy focused on prediction of the user's performance from the site and comparison between expected and actual Web site usage, change of the Web site to the benefits of its operators. From the process of data research of Web usage mining the Web content and Web site topology will be used as the facts sources which relates Web usage mining with the Web content mining and Web structure mining.

Nowadays Web usage mining is selective new research area and expansion further considerations in recent years.

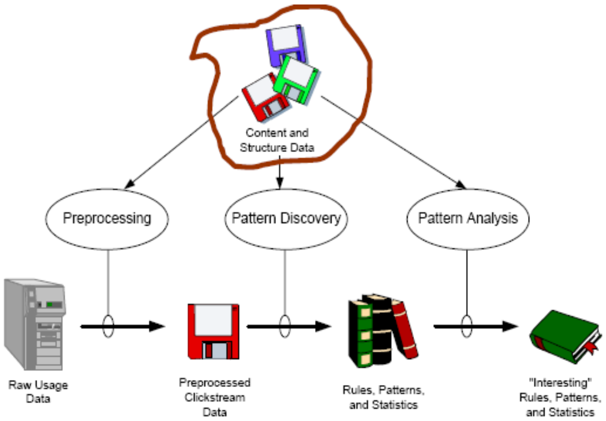

Content and Structure Data
Preprocessing
Pattern Discovery
Pattern Analysis
Raw Usage Data
Preprocessed Clickstream Data
Rules, Patterns, and Statistics
"Interesting" Rules, Patterns, and Statistics

Fig. 1 Web usage mining process.

A. Data Preprocessing for mining

As per consideration of technical point of view preprocessing of data is focus as application of data mining. The purpose of analysed data for web site designing, web personalization, site modification, business intelligence [4]. Before applying any data mining technique on access log data should be well arranged and structuralized form. Preprocessing though work manage data in structuerized and meaningful way through result of any mining technique is unambiguous and Error free. Since the facts construct is very significant in the data preprocess it is necessary to explain the meanings of the connected data notions previously the explanation of the unlike type of the data changing.

- ***User*** –The primary using a visitor to interactively recover and extract resources or resource appearances.
- ***Page view***- Graphical version of a Web page in a detailed client atmosphere at a precise point in time. Click stream – A consecutive series of page view demand
- ***User Session*** - A restricted set of user clicks through one or more Web servers
- ***Server Session*** - A group of user clicks to a single Web site during a user session. Also named a visit.
- Episode **-** A subgroup of related user clicks that occur within a user session

USAGE PREPROCESSING

The inputs of the preprocessing section might embrace the web server logs, referral logs, registration files, index server logs, and optionally usage statistics from a previous analysis.

The outputs area unit the user session file, computer file, website topology, and page classifications. It's forever necessary to adopt a knowledge improvement techniques to eliminate the impact of the orthogonal things to the analysis result. The usage preprocessing in all probability is that the toughest task within the internet Usage Mining process thanks to the wholeness of the obtainable information [5]. While not enough information, it's terribly tough to spot the users. The simplest thanks to improve the info quality is to urge user cooperation, however it's demanding in any respect. There exists a conflict between the analysis wants of the analysts (Who need a lot of elaborate usage information collected) And also the privacy wants of the individual users (who need as very little information collected as possible) [3]. However, the heuristics and statistics strategies is familiar & improve the standard of the web usage information [14]. We have a tendency to might notice some approach to undo the matter, however it's not possible to avoid the misidentification fully, since the net is consequently dynamic and versatile [5].

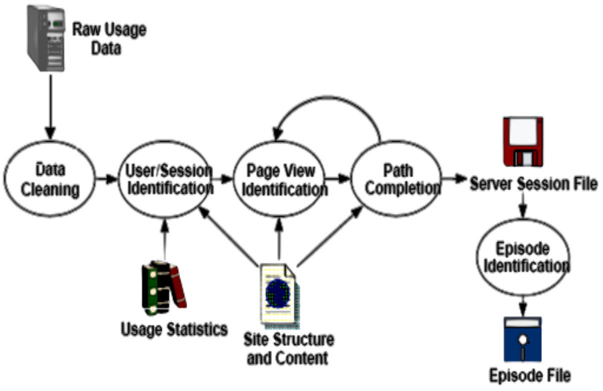


Fig.2 Usage Preprocessing

The session identification is additionally a locality of the usage preprocessing. The goal of it's to divide the page accesses of every user, activity is probably going to go to the net website over once, into individual sessions [19]. The best thanks to do is to use a timeout to interrupt a user's click-stream into session. The thirty minutes is employed as a default timeout by several business merchandise. Another downside is known as as path completion, that indicates the decisive if there are a unit any vital accesses incomprehensible within the access log [2]. Client- or proxy-side caching will usually end in missing access references to those pages or objects that are cached. For example, if a user returns to a page A throughout an equivalent session, the second access to A can probably end in viewing the antecedent downloaded version of A that was cached on the client-side, and so, no request is created to the server. This ends up in the second respect to a not being recorded on the server logs.

III. RELATED WORK

Work first introduce some related effort in data preprocessing and then we focus on need of data warehouse to store the relevant data from the log files created by the web server.

In the recent years, there has been much research on Web usage mining [9], [10], [11], [12]. But then also data preprocessing research field got less attention compare to the other filed of it while actually it deserve to be more attention. Various methods for user identification, session zing, and path completion, are presented already described very well [13].

In another work [14] the authors compared time-based and referrer-based heuristics for visit reconstruction. They found out that a heuristic's appropriateness depends on the design of the Web site such as like whether the site is frame-based or frame-free and depend on the length of the visits because for referrer-based heuristic works better for short visits.

In [15] addressed the application of WUM in the e-learning area with a focus on the preprocessing phase. In this context, they redefined the notion of visit from the e-learning point of view. Moreover, in the same paper, the authors have presented several data preparation techniques to identify Web users, i.e., the path completion and the use of site topology. To identify the user sessions, a fixed time period, say unravel the matter, however it's not possible to avoid the misidentification fully, since the web is therefore dynamic and versatile[5] thirty minutes [6], is used to be the beginning between two sessions. Zaïane et.al[16] have applied various traditional data mining techniques to Internet log files in order to find different types of patterns, which can be harnessed as e-commerce decision support knowledge. The pre-processed data is then loaded into a data warehouse which has an *n*-dimensional web log cube as basis. From this cube, various standard OLAP techniques are applied, such as drilldown, roll-up, slicing, and dicing.

The authors [17] employ the data warehousing technology as a preprocessing step to apply piecewise regression as a predictive data mining technique that fits a data model which will be used for prediction.

A. User Identification

Web user identification is one of the most challenging steps in the process of web usage mining. In case of simple market basket analysis, the customer is identified exactly by its customer ID. However, in case of web users, it is not trivial which page downloads belong to which user. The same individual can use multiple computers, and more persons can use the same computer as well. Furthermore, proxy servers can hide relevant information about unique users as multiple computers appear on the internet using the same IP address through the proxy server.

For 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 [11, 12]. 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. For example in case of selecting frequently visited pages for server side caching, or preloading the next page of common navigational paths, it is irrelevant, whether a page is visited by two different individuals or by one individual twice. The key point is that the page is visited twice. However, in case of an advertisement, it is important, whether two unique individual has seen the page or not.

In many cases other heuristics approached are used for improve identification of the users. In [13] the different 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 behaviour [1]. For example in [14] web users are distinguished based on their navigational patterns using clustering methods.

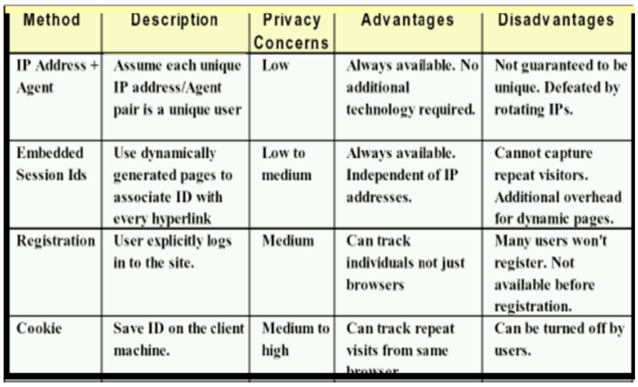
B. Problem at time of User identification

Lots of persons interact everyday with web sites around the world. Massive amount of data are being generated and these information could be much respected to the company in the field of accepting Client's behaviours. Web usage mining is relative independent, but not sequestered category, which mainly describes the techniques that discover the user's usage pattern and try to predict the user's behaviours. Web usage mining is the area of data mining which deals with the novelty and study of usage patterns with use of Web log data. Specifically web logs in direction to advance web based applications.

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 users. In fact, there are masses of users do not register their information. In fact there are great numbers of users access Web sites through, agent, numerous users use the same computer, firewall's existence, independent user use different browsers, and so forth. All of difficulties mark this job greatly complicated and very tough, to identify every unique user accurately. We may use cookies to track users' behaviours. But considering somebody privacy, many users do not practice cookies, so it is needed to find other methods to solve this problem. For users who use the similar computer or use the similar agent, how to find them?

As presented in [9], it uses heuristic method to solve the problem, if a page is requested that is not directly reachable by a hyperlink with some of the, pages visited by the user, the experiential assumes that there is another user with the equal computer or with the equal IPaddress. Doru Tanasa and Brigitte Trousse [4] presents a method called navigation designs to recognize users automatically. But use with this results are not accurate because they only consider a few parts that impact the procedure of user's identification. Rushing to analyse usage data without a proper preprocessing method will lead to poor results or even to failure. This was the case for some of the first WUM tools that were designed to directly extract relationship rules or sequential patterns from the list of Web resources logged for one IP address. Without properly cleaning, transforming and structuring the data prior to the analysis, one cannot expect to find meaningful knowledge. In a KDD process, the preprocessing step represents at least 60% of the entire process for about two thirds of the Data Mining experts responding to the survey1.

Very often, the WUM methods just "translate" the logs into the proper input format for the DM algorithm. However, this would negatively affect the entire process. As an example, let us consider the following hypothetic situation: a Web access log with two requests one for http://website/A/page.htm and another one for http://website/ A/B/C/../../page.htm. The two requests are in fact for the same resource ("../" is interpreted as the parent folder), but they would be treated as two distinct URLs, by most of the translation tools. Therefore, the first problem we aimed to solve was the lack of a complete methodology for preprocessing in WUM.



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.

Fig. 3 Comparison of different user identification

Data Cleaning is done to remove the invalid records with ineffective status, secondary entries with image files, machine navigation entries [15]. Users are recognised to examine user behaviour. In the log files authenticated material is existing for registered websites. In most of the cases these fields are blank due to user's reluctance to use those particular web sites. Time of using any websites Cookies from client side are used for identification but it is not always promising since users might restrict cookies for confidentiality concern. That's why fields used to recognise users are IPaddress, agent (Operating system and Browser) and website topology is also checked to identify a new user by the use of particular links. In time of accessing web pages the requested page is not reachable through any of the pages visited by the user then the user is recognised as a new user in the same IP address [11]. There is rare techniques are existing for user identification with use of different factors like use of http login, Date, Time, client type, cookies info, referrer pages log etc. every technique has focus particular factor of analysis. Through we are assuming result as per fixed condition. It is not sufficient for complete preprocessing result for forthcoming analysis. Work required still more accuracy regarding preprocessing steps.

C. Proposed method of Distinct User Identification

User identification a significant issue is how precisely the users have to be distinguished. It depends principally on the task for the mining procedure is executed. In assured cases the users are identified only with their IPaddress [6]. This can offer an satisfactory result for small time periods like minutes or hours or when the predictable results from the data mining task do not necessity more exactly information about the unique web users. For example in case of selecting regularly visited pages for server side storing or preloading the next page of common directional paths.

In additional cases some heuristics are used for enhanced identification of the users. By [7][6] the different approaches are gathered into two classes the one is the class of the practical methods and the other is that of the responsive methods. Practical strategies aim at distinguishing the users before or through the page request while responsive approaches effort to associate persons with the log entries after the log is written. Practical strategies can be simple user verification with procedures using cookies or using active web pages that are related with the browser likeable them. Responsive strategies work with the verified log files only and the unlike users will be notable by their directional patterns, copy timing sequence or some other heuristics based on some guess regarding their behaviour. For example by [8][6] web users are distinguished grounded on their directional patterns using Time based heuristics technique in which work find how much time user spent with particular page based on them interest. Work can find interested user list from their access Directional strategy.

Considering this actuality, we presented a new algorithm called **"DUI (DISTINCT USER IDENTIFICATION)"**. It studies more factors, such as user's IP address, Web site's topology, browser's edition, operating system, referrer page, and staring time and ending time, requested page status etc. With this algorithm present desirable accuracy and expansibility. It can identify unique users with respect to their session time by total time spent by user with particular page heuristics technique. Work 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.

Definition: given a clean and filtered web log file and record set web log file

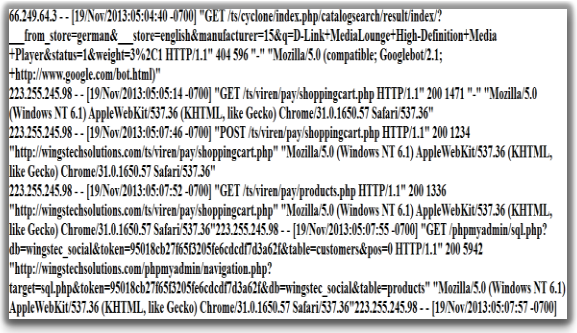
Records Rec= {rec1, rec2, rec3……rec.n} where nrec > 0

```
Line_1: Input Log database RrcUser of N records
Line_2: Distinct User identification base
Line_3:RecUser=P<url, IP-addr, IP-Agent, IP-Method, IP-OS, IP-Status,
        Session-id, IP-Start-Time, IP-End-Time, Total-time-stamp>
// if (isset($_REQUEST['start_date']) && $_REQUEST['end_date'] !='')
{
        $search = " and ac_time >= '".$_REQUEST['start_date']." 00:00:00' and ac_time <= '".$_REQUEST['end_date']." 00:00:00' ";
}
else
{       $search = ""; //

Line_4: RrcUSer=<rec1,rrc2,rec3…recn> where nrec!=0,i=0
Line_5: while(I < nrec)
Line_6: read Logdatabase RecUser
Line_7 check if Rec (i).IP-addr not part of Distinct user identification base then it treated as new user and copy IP-addr in distinct user identification base.
Line_8: end if
Line_9: i=i+1;
Line_10:end loop
Line_11:end
```

IV. IMPLEMENTATION AND RESULT ANALYSIS

Server Log File is the input for our experimental work of preprocessing block. A Web server log is a file to which the Web server writes facts each time a user requests a resource from that particular site. Data preprocessing a web usage mining model aims to reformat the original web logs to identify user's access terms. The Web server usually registers all users' access activities of the website as Web server logs. The data set available in http://wingstechsolutions.com/ consists of a random sample of users visiting the site for a one day period. The Server log file has been used for experimental purposes. It contains unfiltered, unmanaged and rough preprocessed, sessionized data. To effectiveness and efficiency of our methodology mentioned above, with valid we have to use web server logs. November 19, 2013 Initial data source for our experiment which size is around 69.84 MB. After data cleaning, the number of requests declined from 4875 to 3508 remaining all the rows which are not compulsory for web usage mining process was removed. Initially our experimental log file which unfiltered and unmanaged so in fig.4



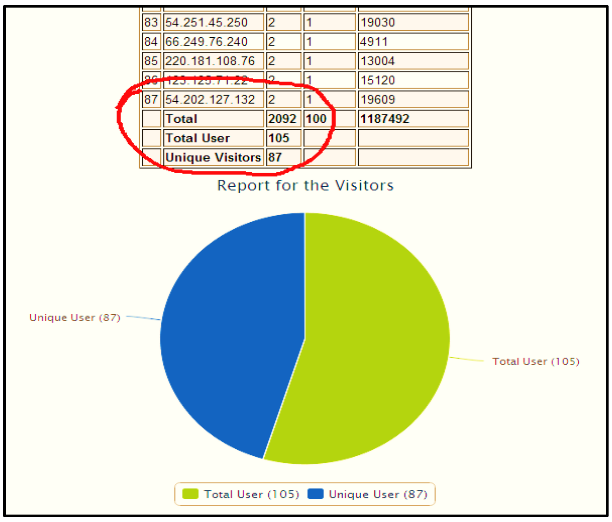
```
66.249.64.3 - - [19/Nov/2013:05:04:40 -0700] "GET /ts/cyclone/index.php/catalogsearch/result/index/?
__from_store=german&__store=english&manufacturer=15&q=D-Link+MediaLounge+High-Definition+Media
+Player&status=1&weight=3%2C1 HTTP/1.1" 404 596 "-" "Mozilla/5.0 (compatible; Googlebot/2.1;
+http://www.google.com/bot.html)"
223.255.245.98 - - [19/Nov/2013:05:05:14 -0700] "GET /ts/viren/pay/shoppingcart.php HTTP/1.1" 200 1471 "-" "Mozilla/5.0
(Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/31.0.1650.57 Safari/537.36"
223.255.245.98 - - [19/Nov/2013:05:07:46 -0700] "POST /ts/viren/pay/shoppingcart.php HTTP/1.1" 200 1234
"http://wingstechsolutions.com/ts/viren/pay/shoppingcart.php" "Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML,
like Gecko) Chrome/31.0.1650.57 Safari/537.36"
223.255.245.98 - - [19/Nov/2013:05:07:52 -0700] "GET /ts/viren/pay/products.php HTTP/1.1" 200 1336
"http://wingstechsolutions.com/ts/viren/pay/shoppingcart.php" "Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML,
like Gecko) Chrome/31.0.1650.57 Safari/537.36"223.255.245.98 - - [19/Nov/2013:05:07:55 -0700] "GET /phpmyadmin/sql.php?
db=wingstec_social&token=95018cb27f65f3205fe6cdcdf7d3a62f&table=customers&pos=0 HTTP/1.1" 200 5942
"http://wingstechsolutions.com/phpmyadmin/navigation.php?
target=sql.php&token=95018cb27f65f3205fe6cdcdf7d3a62f&db=wingstec_social&table=products" "Mozilla/5.0 (Windows NT 6.1)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/31.0.1650.57 Safari/537.36"223.255.245.98 - - [19/Nov/2013:05:07:57 -0700]
```

Fig. 4- Read access log file from server

Before cleaning process start field extraction is required due to focused on particular field for removing irrelevant information for user identification performance. There is no. of files are created with respected to user requested page like when any user click on any page with reference linkup pages are also be operated by user which is not requested by particular user. So work is not consider that types of requested page as user interest. Work need to remove that type of all the request before proceed for user identification. If work want to check individual access of each user who has same ip address through press click bottom. Result get completed information of web pages which is request done by same IPaddress. Each information represent request time and response time with their status code. Through this result work got current status of requested page like 200 to 300 which represent correct response but status like 400 represent error message in a situation server down and page cannot found. Data cleaning is first step of Preprocessing in our experiment on one day server log work can differentiate before data cleaning process work got 4875 rows and after applying cleaning process on one day server log work got 3508 rows. Work reduce 1367 rows which are irrelevant and unnecessary data for feather mining processes through work improve quality of analysis.

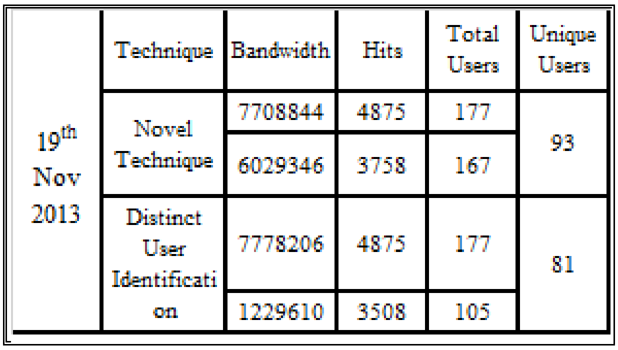
The input for these experiments was output of the data cleaning task for usage preprocessing removing all graphics requests, and removing all log entries that did not contain an fixed period user attention. This included removing a number of hidden POST requests, since the fixed user attended session time was not recorded in the Server log. Each distinct URL for both the request and referrer field was given a numeric named.

Basic User Identification technique which describe in Chapter 4. Latest one for user identification is Novel Technique with respected to IP address and Agent parameter of user access log. Result of Novel technique got weak performance as compare to DUI. Where DUI focused on IPAddress, Agent, Status code, Referer pages, Start_time, End_time factors of server log entry. Through work got complete information of user interest how much time user spent with particular page.



83	54.251.45.250	2	1	19030
84	66.249.76.240	2	1	4911
85	220.181.108.76	2	1	13004
86	123.125.71.22	2	1	15120
87	54.202.127.132	2	1	19609
	Total	**2092**	**100**	**1187492**
	Total User	**105**		
	Unique Visitors	**87**		

Fig. 5- Result analysis of DUI



	Technique	Bandwidth	Hits	Total Users	Unique Users
19th Nov 2013	Novel Technique	7708844	4875	177	93
		6029346	3758	167	
	Distinct User Identification	7778206	4875	177	81
		1229610	3508	105	

Fig. 6 Comparison based on technique

V. CONCLUSION

In this Research Work present Distinct User Identification technique which Expand overall performance of preprocessing steps of web log usage data in Web mining. Work use two Preprocessing Steps combine within one step .Here introduced one proposed algorithm for Unique User Identification is very efficient as compare to other identification techniques. DUI (Distinct User Identification) centred with different factors like IP address, Operating system, chosen browser, Referred pages as well as how much time user spent with particular page on desired session time, requested page Status. It can be used in unauthorized access of data, fake detection, as well as through discovery of regular access process get better designing for future access. Experiments have proved that advanced data preprocessing technique can enhanced the quality of data preprocessing results. As well as work can improve overall performance of entire Preprocessing technique. We get more precious accurate result. Based on this result can easily modified websites, expand the overall design of Webpages.

As usages of users on websites. The new approach of storing preprocessed data using Optimize schema is an increasingly important platform for data examination and online logical processing which will provide an effective platform for data web mining. Future work involves various data alteration tasks that are likely to impact the quality of the discovered Patterns Discovering and Study from the mining algorithms. Future work must be done to mixture whole method of WUM. A whole technology covering like pattern discovery and pattern analysis are additional helpful in identification method.
The exposed patterns can be used for various Web usage applications such as site development, commercial intelligence and recommendations.

REFERENCES

[1] Renáta Iváncsy, and Sándor Juhász, "Analysis of Web User Identification Methods" World Academy of Science, Engineering and Technology 34, pp. 338 to 345, 2007.

[2] V.Chitraa, Dr.Antony Selvadoss Thanamani, "A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing" , International Journal of Computer Applications, ISSN 0975-8887,Volume 34-no 9,November 2011.

[3] Yan LI, Boqin FENG and Qinjiao MAO, "Research on Path Completion Technique in Web Usage Mining", IEEE International Symposium on Computer Science and Computational Technology, pp. 554-559, 2008.

[4] Tasawar Hussain, Dr. Sohail Asghar and Nayyer Masood, "Hierarchical Sessionization at Preprocessing Level of WUM Based on Swarm Intelligence ", 6th International Conference on Emerging Technologies (ICET) IEEE, pp. 21-26, 2010.

[5] Doru Tanasa and Brigitte Trousse, "Advanced Data Preprocessing for Intersites Web Usage Mining ", Published by the IEEE Computer Society, pp. 59-65, March/April 2004.

[6] Huiping Peng, "Discovery of Interesting Association Rules Based On Web Usage Mining", IEEE Coference, pp.272-275, 2010.

[7] Ling Zheng, Hui Gui and Feng Li, " Optimized Data Preprocessing Technology For Web Log Mining", IEEE International Conference On Computer Design and Applications(ICCDA), pp. VI-19-VI-21,2010.

[8] JING Chang-bin and Chen Li, " Web Log Data Preprocessing Based On Collaborative Filtering ", IEEE 2nd International Workshop On Education Technology and Computer Science, pp.118-121, 2010.

[9] Fang Yuankang and Huang Zhiqiu, " A Session Identification Algorithm Based on Frame Page and Pagethreshould", IEEE Conference, pp.645647, 2010.

[10] J. Vellingiri and S. Chenthur Pandian, "A Novel Technique for Web Log Mining with Better Data Cleaning and Transaction Identification", Journal of Computer Science, pp. 683-689, 2011.

[11] Rana Forsati, Mohammad Reza Meybodi and Afsaneh Rahbar, "An Efficient Algorithm for Web Recommendation Systems", IEEE Conference, pp. 579-586, 2009.

[12] D.Vasumathi, D.Vasumathi and K.Suresh, "Effective Web Personalization Using Clustering", IEEE IAMA, 2009.

[13] Zhiguo Zhu and Liping Kong, " A Design For Architecture Model Of Web Access Patterns Mining System ",IEEE International Conference on Computer and Communication Technologies In Agriculture Engineering, pp.288-292, 2010.

[14] Chaoyang Xiang, Shenghui He and Lei Chen, "A Studying System Based On Web Mining", IEEE International Symposium On Intelligent Ubiquitous Computing and Education, pp.433-435, 2009.

[15] R.M.Suresh and Padmajavalli, " An Overview Of Data Preprocessing In Data and Web Usage Mining", IEEE Conference, pp.193-198, 2006.

[16] Margaret H. Dunham, "Data Mining Introductory and Advanced Topics", Pearson Education.

[17] Zaïane, O.R. Xin and M. Han. "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs," In Proceedings of Advances in Digital Libraries Conference (1998), pp. 19-29.

[18] Kamal A. ElDahshan, Hany Maher Said Lala Kamal "Data Warehouse based Statistical Mining," In ICGST-AIML Journal, ISSN: 1687-4846, Volume 9, Issue I, February (2009), pp.41-48

[19] R. Agrawal. Data mining: Crossing the chasm. Invited talk at the 5th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining (KDD99), 1999.