Volume 3, Issue 9, September 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Sessionization Process for the Pages Designed with the Concept of CMS

Nirali Honest, Dr. Atul Patel
Smt. Chandaben Mohanbhai Patel Insitute of Computer Applications, ,
Charotar University of Science and Technology (CHARUSAT),Changa, India

Dr. Bankim Patel
Shrimad Rajchandra Institute of Management and Computer Application
Uka Tarsadia University,
Bardoli, India

Abstract*— Web Usage Mining (WUM) is the process of generating interesting behavior patterns that helps in analyzing of website usage. The frequency access of to any website is not consistent for the given span of interval. The administrator has to mine the patterns for knowing the regular usage and usage during the increased access of the website. In this paper we discuss the problems created with the pages designed with the concept of Content Management System (CMS) and show the mining process for supporting this problem.

Keywords*— Web Usage Mining, Master Page, Data Preprocessing, CMS, Sessionization process by merging uri_stem and uri_query.

I. Introduction

The early methods of WUM can be seen in [1] and [2]. In the recent years, there has been much research on WUM [3], [4], [5], [6]. However, data preprocessing has received far less attention than it deserves. Data Preprocessing is the basic and first step of the Web Usage mining process, the outcomes of this phase impact the next phases as it is an pipelining of input and output from one phase to another. If we get better and accurate result from the first step then only we can improve the mined patterns' quality and save algorithm's running time. It is especially important to process web log files, in respect that the structure of web log files are not the same, they are not structured and complete due to various causations. So it is especially necessary to pre-process web log files in web usage mining. As more organizations make use of the Internet and the World Wide Web to convey information, the conventional approaches to web site evaluation need to be revised [7]. Before the web site can be improved, its current usage needs to be evaluated. Usability evaluation is the process of collecting data about the usability of a design by a specified group of users for a specific activity within a specified environment [8]. The representation of these pages or scripts originates from the web server log. A user session is the click stream of web pages for a single user across the entire web [9].Methods for user identification, sessionzing, and path completion, are presented in [10]. WUM [16] techniques give description about the behaviour of the users , so that we can correlate the relationship of in order to extract relationships in the recorded data.

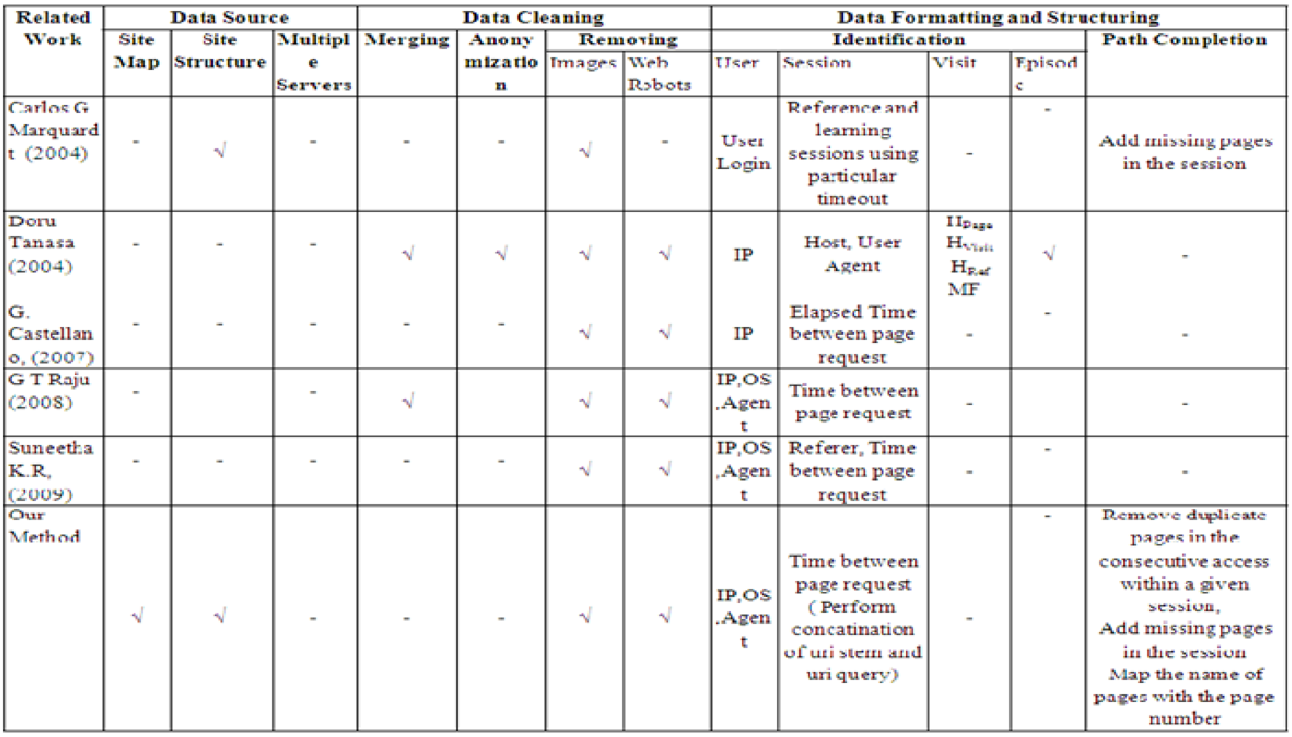
The paper is organized into sections which include significance of Preprocessing phase, Steps taken to clean the log data, identify users, perform sessionization and show the experimental results, followed by conclusion and references.

II. Significance of Preprocessing Phase

Preprocessing phase is an important phase in WUM and it needs to be moulded with different parameters based on the purpose of mining process. The paper shows the mining process for the scenario where a website is designed using CMS in which master page is designed and content of secondary pages is loaded from the database. These secondary pages are given unique numbers to retrieve from the database. Due to these characteristics of pages designed using CMS it is difficult to identify these pages uniquely by name, so this is one of the troubleshooting areas when certain patterns are to be generated as part of the mining process [17]. For example if administrator of the site wishes to generate per page frequency of the pages of a given site, then certain tools don’t support the per page frequency of pages designed using CMS, and the tools that support this functionality generate the report using the page ID number, not by name. So the generated reports give per page frequency but it may not be so useful, as looking to Page ID doesn’t give information which page frequency is shown. This is one of the key aspects to focus in our preprocessing phase. The next aspect is to generate the patterns for certain special events of a given website. The special events can be specific to a given website, if we consider a website for University , the access of site will be different when we have some events like admission process, recruitment process, workshop details, etc. To generate the patterns for these special events we need to specify interval of time during which these events are occurred. This feature is available with current tools, but we need to specify a range of dates for which we need to generate the patterns, for the administrator of site it is not possible to remember all the range of dates for which these patterns are to be generated. This is the second key issues for our mining process. To achieve the above goals of our mining process we have moulded the preprocessing phase according to our requirement. The major purpose of preprocessing phase we are designing is to extract useful data from raw web log and then transform these data in to the form necessary for further processing. This phase leads to the benefits of reducing the size of log file and increase the quality of the data.

The basic steps of our preprocessing phase[18] includes Data cleaning, User Identification, Session Identification, Path Completion, Generating Academic Events and Generating site structure which include generating site map and mapping of page ID with page Name. This paper includes the algorithm and results for the first three steps: Data Cleaning, User Identification and Session Identification. WUM application in the distance education domain is given by Carlos G. Marquardt, [11] in which they mould the phase of preprocessing as their requirement. Doru Tanasa et al. [12] perform data structuration and summarization to form the Data warehouse construction. A tool namely LODAP was constructed by G. Castellano et al.[13] to perform data cleaning, data structuration and data filtering. K. R. Suneetha et al.[[14] performed Data Cleaning, User Identification , Session Identification and constructed data warehouse. G T Raju, et. al.[15] performed Merging of log files, Data Cleaning, Identification of Users, Sessions, Page Views, and Visits. Table 1 shows the comparison of the proposed work with the related work.

TABLE 1
COMPARISON of the Proposed work with the Related work.



Related Work	Data Source			Data Cleaning				Data Formatting and Structuring				
	Site Map	Site Structure	Multiple Servers	Merging	Anonymization	Removing		Identification				Path Completion
						Images	Web Robots	User	Session	Visit	Episode	
Carlos G Marquardt (2004)	-	√	-	-	-	√	-	User Login	Reference and learning sessions using particular timeout	-	-	Add missing pages in the session
Doru Tanasa (2004)	-	-	-	√	√	√	√	IP	Host, User Agent	H_{Page} H_{Visit} H_{Ref} MF	√	-
G. Castellano, (2007)	-	-	-	-	-	√	√	IP	Elapsed Time between page request	-	-	-
G T Raju (2008)	-		-	√		√	√	IP,OS ,Agent	Time between page request	-		-
Suneetha K.R, (2009)	-	-	-	-	-	√	√	IP,OS ,Agent	Referer, Time between page request	-	-	-
Our Method	√	√	-	-	-	√	√	IP,OS ,Agent	Time between page request (Perform concatination of uri stem and uri query)	-	-	Remove duplicate pages in the consecutive access within a given session, Add missing pages in the session Map the name of pages with the page number

III. DATA CLEANING AND USER IDENTIFICATION

Data Cleaning : Data cleaning means eliminate the irrelevant information (information which is not useful for the further process, or even misleading or faulty)from the original Web log file and exchange the Web log as data-base which is convenient for further processing. Steps for Data Cleaning are as below,

- Download the W3C Extended Log file from internet.
- Parse the raw log file according to delimiter and convert it to appropriate fields of W3C Extended log file format.
- Remove all other entries which have other then .html, .asp,.aspx,.php extensions. These also include log entries which do not have any URL in the URL entry.
- Remove log entries having code other then 200 and 304 from the file.
- Remove entries with request methods except GET and POST.
- Remove web crawlers, robots, Spiders.

After the implementation of these steps cleaned log data is stored in the database.

User Identification: User identification means identifying each user accessing Web site, whose goal is to mine every user's access characteristic. Steps for User Identification

- Read record from the cleaned log database.
- If new IP address then add new record the IP address, browser and OS details and increment the count of number of users.
- If IP address is already present then compare the browser and OS details if not same then increment the count of number of users.
- Repeat the above steps until all records are processed.

After the implementation of these steps user are identified and the data is stored in the database.

IV. Session Identification

After user identification, the pages accessed by each user must be divided into individual sessions, which is known as session identification. The goal of session identification is to find what users accessed during the visit to the website and correlate multiple requests from each user. There are various problems that make this phase complex, some of them are HTTP is a stateless protocol which means that every request is taken independent of the other request, there is no relation in the multiple requests given by the same client, Catching is performed either by proxy server or browser, which means that a single IP address can be associated with different user sessions, User may visit the website more than once so the server logs more than one entry for the user. Various session identification techniques have been proposed like Time based Heuristics using combination of IP address and User Agent, Time based Heuristics using Cookies, User explicitly logs in by registration. This proposed research uses first technique mentioned above, IP address + User-Agent + Time Heuristics to generate sessions. The main benefit of considering this technique is, that the details are always available as part of log file and no extra detailing is required to perform sessionization.

The simplest method of achieving this is through a timeout, where if the time between page requests exceeds a certain limit, it is assumed that the user is starting a new session. We use 30 minutes as a default timeout based on the empirical data.

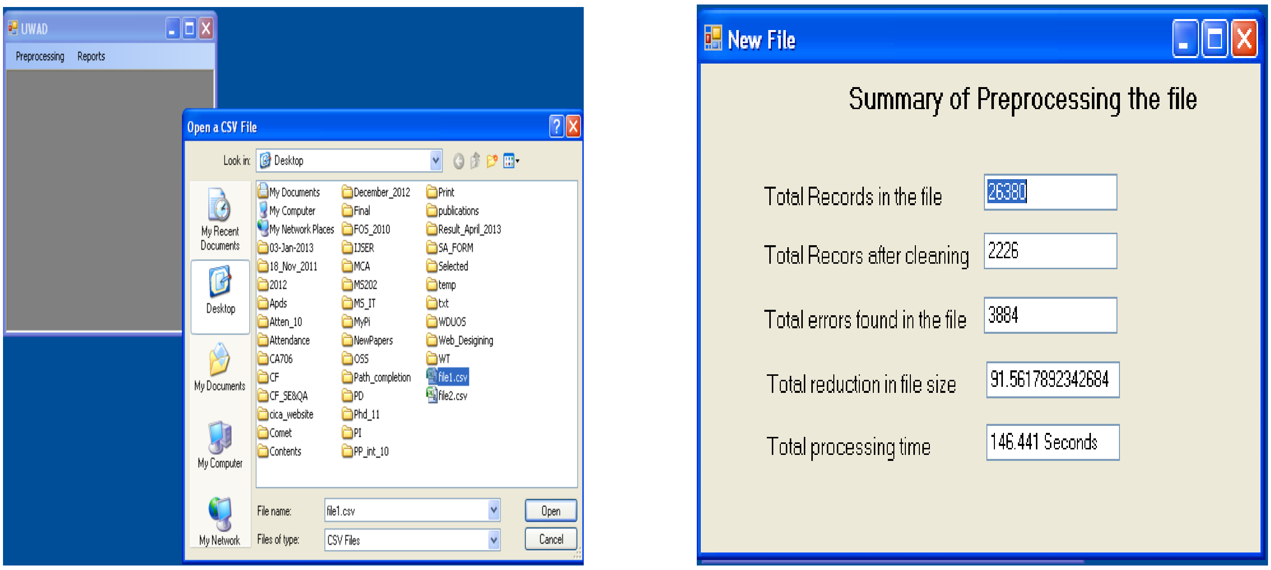
Steps of session identification

- Read records from the log file of a particular date.
- If there is a new user, then there is a new session.
- If the time between the page requests exceeds a certain limits (30 Minutes), it is assumed that the user is starting a new session. (For same IP addresses)
- Read the uri_stem & uri_query for each record and concate it, to form the uri.
- Repeat the above steps until all sessions are formed
- Once sessions are identified, read individual sessions for a given user concat all the pages accessed by the users in a given session and time taken to access all pages in a given session.

After the implementation of these steps sessions for the users are identified and the data is stored in the database.

V. Experimental Results

Currently there are lots of open source tools available to perform pre-processing task, for example log parser lizard is one of the open source tools available for performing the required task. Initially we had implemented the algorithm steps using Log parser lizard tool [18], but for performing steps with the tool we need to perform lots of intermediate tasks like, managing to write queries, storing intermediate results and performing other utility tasks using MS Excel, for cleaning a single file. All these activities include lots of extra overhead so it becomes a error prone and time consuming process. Also certain automated facilities are not as per the requirement of the moulded mining process. Another objective is to generate per page frequency but the pages designed with CMS are generated with page ID and not page Name, so biding of ID and Name is not done by certain tools like state counter is a tool which support the generation of per page frequency but it generates report with page ID and not with name so it may not be informative. So taking these points into consideration we have prepared our own tool for our mining process namely UWAD (University Website Access Domain), and now the above algorithm steps are implemented using this tool. The next session shows the results generated using this tool. Figure 1 shows the use of tool for selecting the file to pre-process and after that the summary of details of the cleaning process. Table3 shows the summary of the Data Cleaning process. Figure 2 reflects the cleaned log data.



Look in: Desktop

File name: file1.csv

Files of type: CSV Files

Summary of Preprocessing the file

Total Records in the file: 26380

Total Recors after cleaning: 2226

Total errors found in the file: 3884

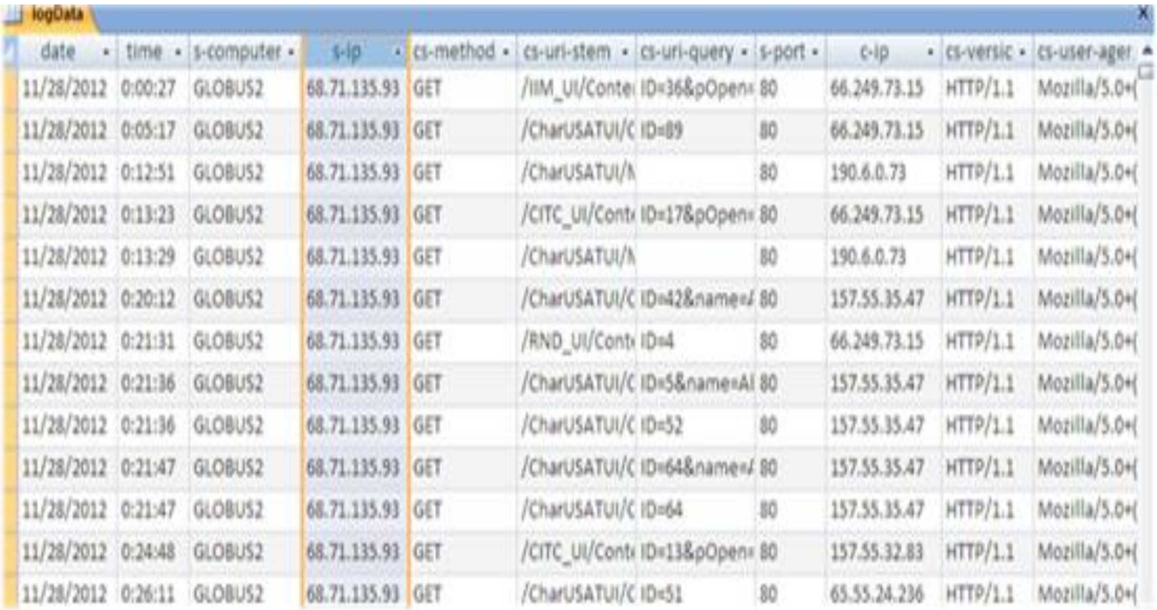
Total reduction in file size: 91.5617892342684

Total processing time: 146.441 Seconds

Fig. 1 (a) : Shows the tool to select file for preprocessing (b) : Shows the summary of the cleaning process

TABLE 2
SUMMARY OF THE CLEANING PROCESS.

Stage of Preprocessing	No. of Web Objects Retrieved
Total records in the file	26380
Total records after cleaning	2226
Total errors found in the file	3884
Total reduction in file size	91.56%
Total time taken for data cleaning	146.441 Seconds (2.44 minutes)

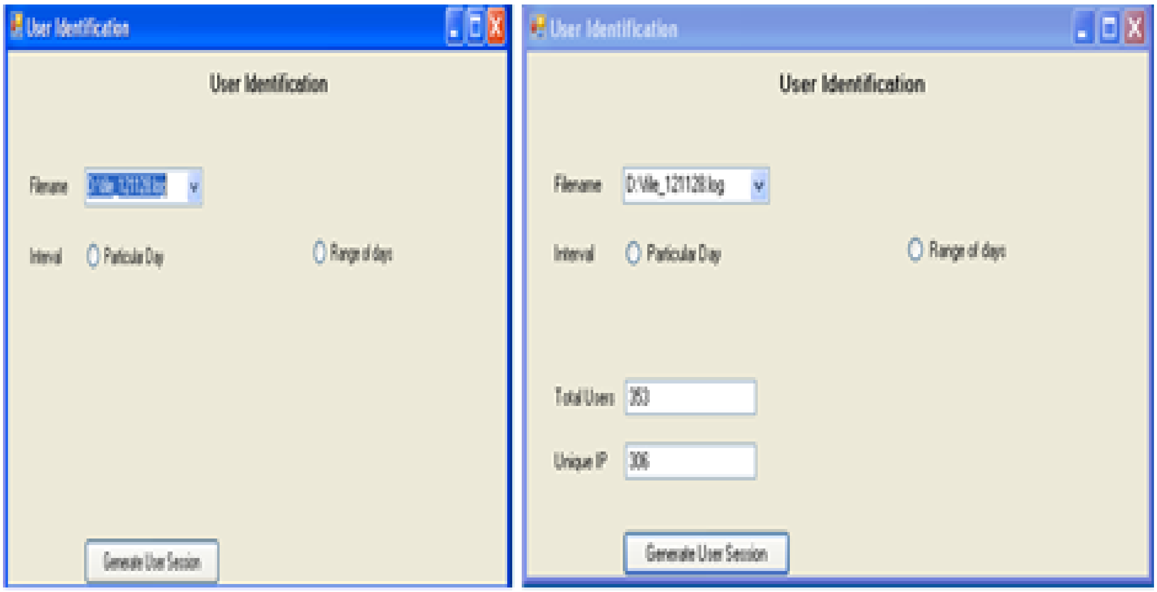


logData

date	time	s-computer	s-ip	cs-method	cs-uri-stem	cs-uri-query	s-port	c-ip	cs-versic	cs-user-ager
11/28/2012	0:00:27	GLOBUS2	68.71.135.93	GET	/IIM_UI/Conte	ID=36&pOpen=	80	66.249.73.15	HTTP/1.1	Mozilla/5.0+(
11/28/2012	0:05:17	GLOBUS2	68.71.135.93	GET	/CharUSATUI/C	ID=89	80	66.249.73.15	HTTP/1.1	Mozilla/5.0+(
11/28/2012	0:12:51	GLOBUS2	68.71.135.93	GET	/CharUSATUI/N		80	190.6.0.73	HTTP/1.1	Mozilla/5.0+(
11/28/2012	0:13:23	GLOBUS2	68.71.135.93	GET	/CITC_UI/Conte	ID=17&pOpen=	80	66.249.73.15	HTTP/1.1	Mozilla/5.0+(
11/28/2012	0:13:29	GLOBUS2	68.71.135.93	GET	/CharUSATUI/N		80	190.6.0.73	HTTP/1.1	Mozilla/5.0+(
11/28/2012	0:20:12	GLOBUS2	68.71.135.93	GET	/CharUSATUI/C	ID=42&name=/	80	157.55.35.47	HTTP/1.1	Mozilla/5.0+(
11/28/2012	0:21:31	GLOBUS2	68.71.135.93	GET	/RND_UI/Conte	ID=4	80	66.249.73.15	HTTP/1.1	Mozilla/5.0+(
11/28/2012	0:21:36	GLOBUS2	68.71.135.93	GET	/CharUSATUI/C	ID=5&name=Al	80	157.55.35.47	HTTP/1.1	Mozilla/5.0+(
11/28/2012	0:21:36	GLOBUS2	68.71.135.93	GET	/CharUSATUI/C	ID=52	80	157.55.35.47	HTTP/1.1	Mozilla/5.0+(
11/28/2012	0:21:47	GLOBUS2	68.71.135.93	GET	/CharUSATUI/C	ID=64&name=/	80	157.55.35.47	HTTP/1.1	Mozilla/5.0+(
11/28/2012	0:21:47	GLOBUS2	68.71.135.93	GET	/CharUSATUI/C	ID=64	80	157.55.35.47	HTTP/1.1	Mozilla/5.0+(
11/28/2012	0:24:48	GLOBUS2	68.71.135.93	GET	/CITC_UI/Conte	ID=13&pOpen=	80	157.55.32.83	HTTP/1.1	Mozilla/5.0+(
11/28/2012	0:26:11	GLOBUS2	68.71.135.93	GET	/CharUSATUI/C	ID=51	80	65.55.24.236	HTTP/1.1	Mozilla/5.0+(

Fig. 2 Snapshot of log data after the cleaning process

Figure 3 shows the interface for selecting the cleaned log file for User Identification and shows the total users and unique users of the file. Figure 4 show the entry of users generated from the file. Figure 5 shows that a user with same IP can be treated as a different user as the agent is different.



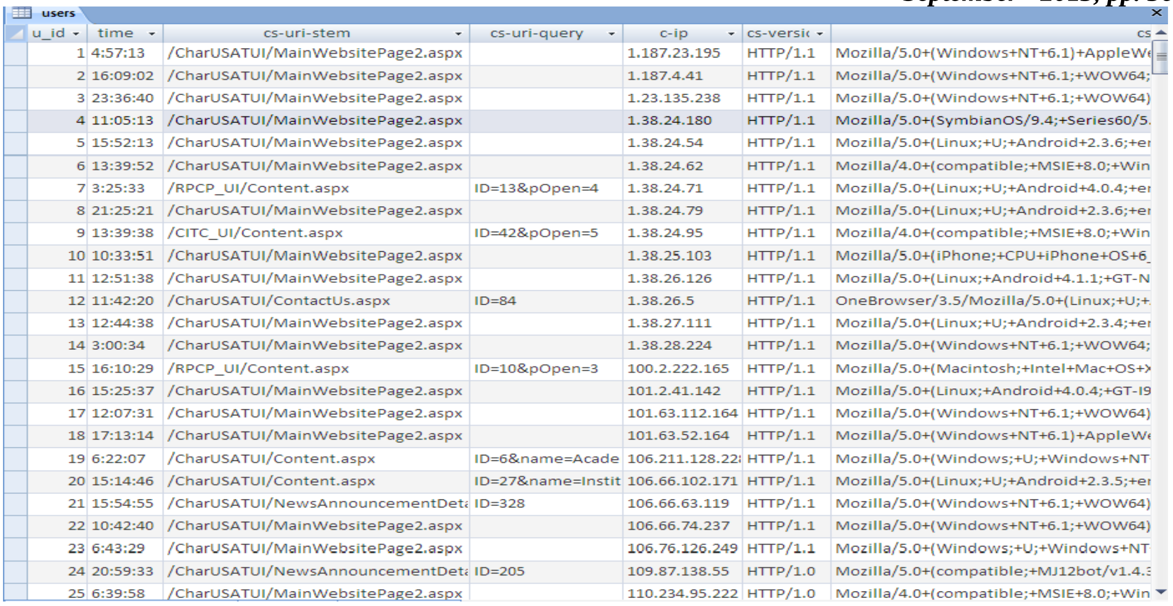
Filename D:\file_121128.log

Filename D:\file_121128.log

Total Users 353

Unique IP 306

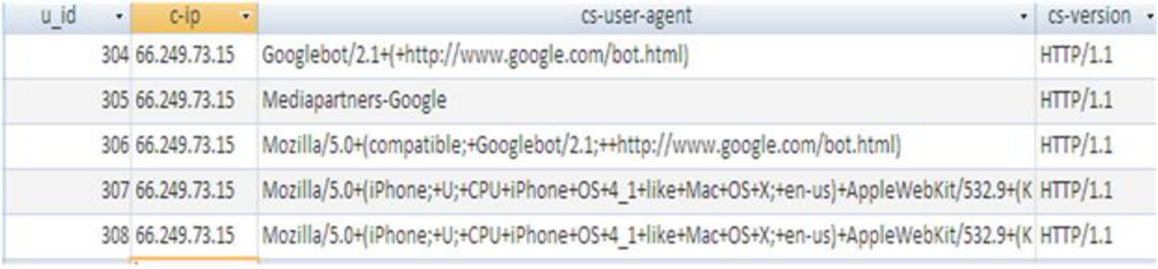
Fig. 3 (a) Selecting Cleaned log for User Identification (b) Generating Total Users and Unique Users



users

u_id	time	cs-uri-stem	cs-uri-query	c-ip	cs-versi	cs
1	4:57:13	/CharUSATUI/MainWebsitePage2.aspx		1.187.23.195	HTTP/1.1	Mozilla/5.0+(Windows+NT+6.1)+AppleWe
2	16:09:02	/CharUSATUI/MainWebsitePage2.aspx		1.187.4.41	HTTP/1.1	Mozilla/5.0+(Windows+NT+6.1;+WOW64;
3	23:36:40	/CharUSATUI/MainWebsitePage2.aspx		1.23.135.238	HTTP/1.1	Mozilla/5.0+(Windows+NT+6.1;+WOW64)
4	11:05:13	/CharUSATUI/MainWebsitePage2.aspx		1.38.24.180	HTTP/1.1	Mozilla/5.0+(SymbianOS/9.4;+Series60/5.
5	15:52:13	/CharUSATUI/MainWebsitePage2.aspx		1.38.24.54	HTTP/1.1	Mozilla/5.0+(Linux;+U;+Android+2.3.6;+er
6	13:39:52	/CharUSATUI/MainWebsitePage2.aspx		1.38.24.62	HTTP/1.1	Mozilla/4.0+(compatible;+MSIE+8.0;+Win
7	3:25:33	/RPCP_UI/Content.aspx	ID=13&pOpen=4	1.38.24.71	HTTP/1.1	Mozilla/5.0+(Linux;+U;+Android+4.0.4;+er
8	21:25:21	/CharUSATUI/MainWebsitePage2.aspx		1.38.24.79	HTTP/1.1	Mozilla/5.0+(Linux;+U;+Android+2.3.6;+er
9	13:39:38	/CITC_UI/Content.aspx	ID=42&pOpen=5	1.38.24.95	HTTP/1.1	Mozilla/4.0+(compatible;+MSIE+8.0;+Win
10	10:33:51	/CharUSATUI/MainWebsitePage2.aspx		1.38.25.103	HTTP/1.1	Mozilla/5.0+(iPhone;+CPU+iPhone+OS+6_
11	12:51:38	/CharUSATUI/MainWebsitePage2.aspx		1.38.26.126	HTTP/1.1	Mozilla/5.0+(Linux;+Android+4.1.1;+GT-N
12	11:42:20	/CharUSATUI/ContactUs.aspx	ID=84	1.38.26.5	HTTP/1.1	OneBrowser/3.5/Mozilla/5.0+(Linux;+U;+
13	12:44:38	/CharUSATUI/MainWebsitePage2.aspx		1.38.27.111	HTTP/1.1	Mozilla/5.0+(Linux;+U;+Android+2.3.4;+er
14	3:00:34	/CharUSATUI/MainWebsitePage2.aspx		1.38.28.224	HTTP/1.1	Mozilla/5.0+(Windows+NT+6.1;+WOW64;
15	16:10:29	/RPCP_UI/Content.aspx	ID=10&pOpen=3	100.2.222.165	HTTP/1.1	Mozilla/5.0+(Macintosh;+Intel+Mac+OS+X
16	15:25:37	/CharUSATUI/MainWebsitePage2.aspx		101.2.41.142	HTTP/1.1	Mozilla/5.0+(Linux;+Android+4.0.4;+GT-I9
17	12:07:31	/CharUSATUI/MainWebsitePage2.aspx		101.63.112.164	HTTP/1.1	Mozilla/5.0+(Windows+NT+6.1;+WOW64)
18	17:13:14	/CharUSATUI/MainWebsitePage2.aspx		101.63.52.164	HTTP/1.1	Mozilla/5.0+(Windows+NT+6.1)+AppleWe
19	6:22:07	/CharUSATUI/Content.aspx	ID=6&name=Acade	106.211.128.22	HTTP/1.1	Mozilla/5.0+(Windows;+U;+Windows+NT
20	15:14:46	/CharUSATUI/Content.aspx	ID=27&name=Instit	106.66.102.171	HTTP/1.1	Mozilla/5.0+(Linux;+U;+Android+2.3.5;+er
21	15:54:55	/CharUSATUI/NewsAnnouncementDet	ID=328	106.66.63.119	HTTP/1.1	Mozilla/5.0+(Windows+NT+6.1;+WOW64)
22	10:42:40	/CharUSATUI/MainWebsitePage2.aspx		106.66.74.237	HTTP/1.1	Mozilla/5.0+(Windows+NT+6.1;+WOW64)
23	6:43:29	/CharUSATUI/MainWebsitePage2.aspx		106.76.126.249	HTTP/1.1	Mozilla/5.0+(Windows;+U;+Windows+NT
24	20:59:33	/CharUSATUI/NewsAnnouncementDet	ID=205	109.87.138.55	HTTP/1.0	Mozilla/5.0+(compatible;+MJ12bot/v1.4.3
25	6:39:58	/CharUSATUI/MainWebsitePage2.aspx		110.234.95.222	HTTP/1.0	Mozilla/4.0+(compatible;+MSIE+8.0;+Win

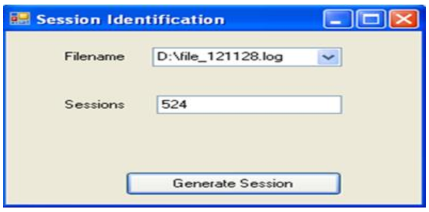
Fig. 4 Snapshot of Users Identified



u_id	c-ip	cs-user-agent	cs-version
304	66.249.73.15	Googlebot/2.1+(+http://www.google.com/bot.html)	HTTP/1.1
305	66.249.73.15	Mediapartners-Google	HTTP/1.1
306	66.249.73.15	Mozilla/5.0+(compatible;+Googlebot/2.1;++http://www.google.com/bot.html)	HTTP/1.1
307	66.249.73.15	Mozilla/5.0+(iPhone;+U;+CPU+iPhone+OS+4_1+like+Mac+OS+X;+en-us)+AppleWebKit/532.9+(K	HTTP/1.1
308	66.249.73.15	Mozilla/5.0+(iPhone;+U;+CPU+iPhone+OS+4_1+like+Mac+OS+X;+en-us)+AppleWebKit/532.9+(K	HTTP/1.1

Fig. 5 Snapshot of 5 different users with same IP address but different user agent

Figure 6 shows the interface for selecting the cleaned log file for Session Identification and shows the total sessions generated for the selected file. Figure 7 show the entry of sessions generated from the file. It shows the individual pages accessed by all the users, in a given session.

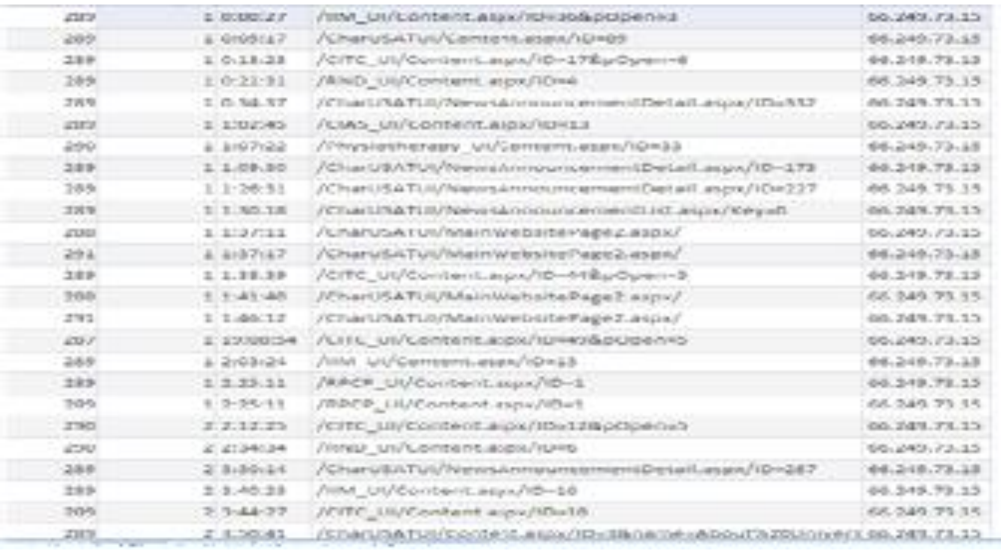


Session Identification

Filename: D:\file_121128.log

Sessions: 524

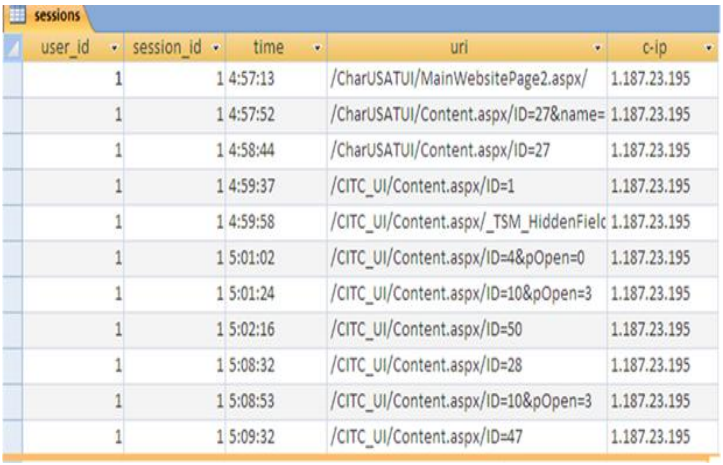
Fig. 6 Snapshot of generating sessions



287	1	0:00:27	/MM_UI/Content.aspx/ID=30&pOpen=3	66.249.73.15
289	1	0:00:17	/CharUSATUI/Content.aspx/ID=89	66.249.73.15
289	1	0:18:28	/CITC_UI/Content.aspx/ID=17&pOpen=6	66.249.73.15
289	1	0:22:31	/AIND_UI/Content.aspx/ID=4	66.249.73.15
289	1	0:34:37	/CharUSATUI/NewsAnnouncementDetail.aspx/ID=337	66.249.73.15
287	1	1:02:40	/CIAS_UI/Content.aspx/ID=13	66.249.73.15
290	1	1:07:22	/Physiotherapy_UI/Content.aspx/ID=33	66.249.73.15
289	1	1:09:30	/CharUSATUI/NewsAnnouncementDetail.aspx/ID=173	66.249.73.15
289	1	1:26:31	/CharUSATUI/NewsAnnouncementDetail.aspx/ID=227	66.249.73.15
289	1	1:30:18	/CharUSATUI/NewsAnnouncementList.aspx/Key=8	66.249.73.15
288	1	1:27:11	/CharUSATUI/MainWebsitePage2.aspx/	66.249.73.15
291	1	1:37:17	/CharUSATUI/MainWebsitePage2.aspx/	66.249.73.15
289	1	1:38:39	/CITC_UI/Content.aspx/ID=41&pOpen=3	66.249.73.15
300	1	1:41:48	/CharUSATUI/MainWebsitePage2.aspx/	66.249.73.15
291	1	1:46:12	/CharUSATUI/MainWebsitePage2.aspx/	66.249.73.15
287	1	1:00:34	/CITC_UI/Content.aspx/ID=43&pOpen=5	66.249.73.15
289	1	2:03:24	/MM_UI/Content.aspx/ID=13	66.249.73.15
289	1	2:23:11	/RPCP_UI/Content.aspx/ID=1	66.249.73.15
305	1	2:25:11	/RPCP_UI/Content.aspx/ID=1	66.249.73.15
290	2	2:12:25	/CITC_UI/Content.aspx/ID=12&pOpen=3	66.249.73.15
290	2	2:34:34	/RNID_UI/Content.aspx/ID=5	66.249.73.15
288	2	3:30:14	/CharUSATUI/NewsAnnouncementDetail.aspx/ID=287	66.249.73.15
289	2	3:40:28	/MM_UI/Content.aspx/ID=16	66.249.73.15
305	2	3:44:27	/CITC_UI/Content.aspx/ID=18	66.249.73.15
287	2	3:30:41	/CharUSATUI/Content.aspx/ID=3&name=About%20Univer	66.249.73.15

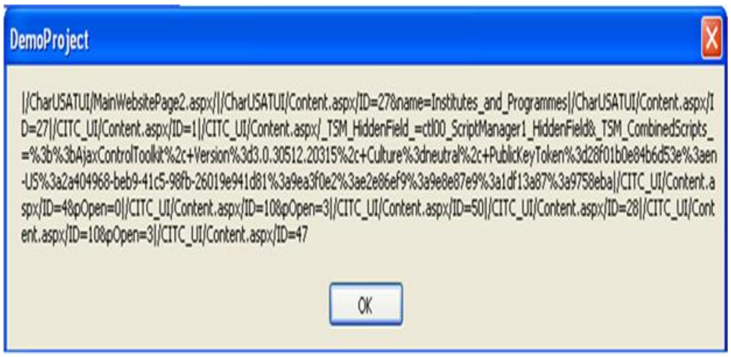
Fig. 7 Snapshot of sessions generated

Figure 8 shows the session generated for an individual user. It shows all the pages accessed by a single user during a given session. After all pages are identified they are merged using a symbol “|”. Figure 9 shows the merging of all pages accessed during a given session for an individual user. Table 4 shows the summary of users and sessions identified for a given file.



user_id	session_id	time	uri	c-ip
1	1	4:57:13	/CharUSATUI/MainWebsitePage2.aspx/	1.187.23.195
1	1	4:57:52	/CharUSATUI/Content.aspx/ID=27&name=	1.187.23.195
1	1	4:58:44	/CharUSATUI/Content.aspx/ID=27	1.187.23.195
1	1	4:59:37	/CITC_UI/Content.aspx/ID=1	1.187.23.195
1	1	4:59:58	/CITC_UI/Content.aspx/_TSM_HiddenFielc	1.187.23.195
1	1	5:01:02	/CITC_UI/Content.aspx/ID=4&pOpen=0	1.187.23.195
1	1	5:01:24	/CITC_UI/Content.aspx/ID=10&pOpen=3	1.187.23.195
1	1	5:02:16	/CITC_UI/Content.aspx/ID=50	1.187.23.195
1	1	5:08:32	/CITC_UI/Content.aspx/ID=28	1.187.23.195
1	1	5:08:53	/CITC_UI/Content.aspx/ID=10&pOpen=3	1.187.23.195
1	1	5:09:32	/CITC_UI/Content.aspx/ID=47	1.187.23.195

Fig. 8 Snapshot of sessions generated for particular user id 1.



```
|/CharUSATUI/MainWebsitePage2.aspx/||/CharUSATUI/Content.aspx/ID=27&name=Institutes_and_Programmes||/CharUSATUI/Content.aspx/I
D=27||/CITC_UI/Content.aspx/ID=1||/CITC_UI/Content.aspx/_TSM_HiddenField_=ctl00_ScriptManager1_HiddenField&_TSM_CombinedScripts_
=%3b%3bAjaxControlToolkit%2c+Version%3d3.0.30512.20315%2c+Culture%3dneutral%2c+PublicKeyToken%3d28f01b0e84b6d53e%3aen
-US%3a2a404968-beb9-41c5-98fb-26019e941d81%3a9ea3f0e2%3ae2e86ef9%3a9e8e87e9%3a1df13a87%3a9758eba||/CITC_UI/Content.a
spx/ID=4&pOpen=0||/CITC_UI/Content.aspx/ID=10&pOpen=3||/CITC_UI/Content.aspx/ID=50||/CITC_UI/Content.aspx/ID=28||/CITC_UI/Cont
ent.aspx/ID=10&pOpen=3||/CITC_UI/Content.aspx/ID=47
```

Fig. 9 Snapshot of all uri’s separated by “|” within a given session for an individual user id 1.

TABLE 3
SUMMARY OF USER AND SESSION IDENTIFICATION

Stage of Preprocessing	No. of Web Objects Retrieved
No. of Entries	2226
No. of Unique IP Addresses	306
No. of Users	353
No. of Sessions	524

VI. CONCLUSION

Performing sessions is an important part of the mining process , we have presented the sessionization process for the pages designed using the CMS, to handle them our process concates the uri_stem & uri_query of individual request which forms the entire uri , and after that all the uri’s given within an identified session are merged to form the total set of pages accessed during a session. This step of mining further leads to input of Path completion where the duplicate pages will be removed and the missing pages will be added followed by mapping of Page ID with Page Name, to make the information more productive.

ACKNOWLEDGMENT

The authors thank Charotar University of Science and Technology (CHARUSAT) for providing necessary resources to accomplish this study.

REFERENCES

[1] H. Mamila, H Toivonen, and A. I. Verkamo, "Discovering Frequent Episodes in Sequences", Proc. of the Is1 lnl. Conj on Knowledge DiseoveryondDoto Mining, Montreal, Canada, August 1995.

[2] Pitkow, "In search of reliable usage data on the WWW", Proc. 6th Int. WWW Gej, Santa Carla, CA, pp. 451463, 1997.

[3] B. Mobasher, H. Dai, T. Luo and M. Nakagawa. " Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization," In proceedings of Data Mining and Knowledge Discovery,2002, pp 61-82.

[4] R. Kosala, H. and Blockeel. "Web mining research: a survey," In proceedings of special interest group on knowledge discovery & data mining, SIGKDD:2000 , ACM 2 (1), pp.1–15.

[5] R. Kohavi and R. Parekh. "Ten supplementary analyses to improve e-commerce web sites," In Proceedings of the Fifth WEBKDD workshop, (2003).

[6] B. Mobasher, R. Cooley and J. Srivastava. "Creating Adaptive Web Sites through usage based clustering of URLs," In proceedings of Knowledge and Data Engineering Exchange, 1999, Volume 1, Issue1, 1999, pp.19-25.

[7] COOLEY, R., MOBASHER, B. and SRIVASTAVA, J. (1999): Web Mining: Information and Pattern Discovery on the World Wide Web. http://www.users.cs.umn.edu/~mobasher/webminer/survey/survey.html

[8] SPILIOPOULOU, M.: Web usage mining for Web site evaluation. Communications of the ACM 43(8):127-134.August,2000.

[9] SRIVASTAVA, J., COOLEY, R., DESHPANDE, M. and TAN, P.-N. (2000): Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. ACM SIGKDDExplorations Newsletter Vol I(2):12-23.January.

[10] J. Srivastava, R. Cooley, M. Deshpande and P. N. Tan. "Web usage mining: discovery and applications of usage patterns from web data," In SIGKDD Explorations, 2002, pp. 12–23.

[11] Carlos G. Marquardt, Karin Becker and Duncan D. Ruiz, " A pre-processing tool for Web Usage Mining in the Distance Education Domain", Proceedings of the International Database Engineering and Applications Symposium (IDEAS'04) IEEE, 2004.

[12] Doru Tanasa and Brigitte Trousse, " Advanced Data Preprocessing for Intersites Web Usage Mining ", IEEE Computer Society, March/April, 2004.

[13] G. Castellano, A. M. Fanelli, M. A. Torsello,"Log data preparation for mining web usage patterns", International Conference Applied Computing (IADIS), 2007.

[14] K. R. Suneetha, Dr. R. Krishnamoorthi, " Identifying User Behavior by Analyzing Web Server Access Log File", International Journal of Computer Science and Network Security, Vol. 9 No. 4, April 2009.

[15] Raju G.T. and Sathyanarayana P. "Knowledge discovery from Web Usage Data : Complete Preprocessing Methodology, ", IJCSNS 2008.

[16] S. Bai ,Q. Han ,Q Liu ,X. Gao "Research of an Algorithm Based on Web Usage Mining" China, IEEE, 2009.

[17] Nirali Honest, Dr. Bankim Patel and Atul Patel. Article"Applying Web Usage Mining to a University Website Access Domain". International Journal of Applied Information Systems 2(9):7-14, June 2012. Published by Foundation of Computer Science, New York, USA.

[18] Nirali Honest, Dr. Bankim Patel and Dr. Atul Patel. Article "Preprocessing phase for University Website Access Domain", International Journal of Scientific & Engineering Research, (IJSER) – ISSN : 2229-5518, 4, No.6, June 2013.