International Journal of Engineering Sciences Paradigms and Researches (IJESPR)
(Vol. 23, Issue 01) and (Publishing Month: August 2015)
(An Indexed, Referred and Impact Factor Journal)
ISSN (Online): 2319-6564
www.ijesonline.com

Data Pre-Processing of Web Server Logs for Mining users Access Patterns

Mohammed Elhebir[1] and Ajith Abraham[2]

[1]Lecturer, Faculty of Mathematical & Computer Sciences, University of Gezira
Wad Medani, Gezira, Sudan
elhibr@uofg.edu.sd

[2]Professor, Machine Intelligence Research Labs (MIR Labs), Scientific Network for Innovation and Research Excellence, WA, USA
ajith.abraham@ieee.org

Publishing Date: 5th August 2015

Abstract

Web Usage Mining (WUM) can be defined as the application of data mining techniques to extract the knowledge hidden in the Web log file, such as user access patterns from Web data in order to analyze users' behavioral patterns. However, the data stored in these log files do not present accurately a picture of the users' accesses to Web sites. The Web Usage Mining process goes through three phases: data pre-processing, patterns discovery and pattern analysis. Pre-processing is one of the most important phases in WUM which transforms a log into a set of web user sessions and makes them suitable for analysis. A sample Web log file was collected from the Web server in Sudan University of Science and Technology (SUST).This paper focuses on the pre-processing which reduces size of logs for mining user access patterns.

Keywords: *Web Usage Mining, Web log data, Pre-processing, Access patterns.*

1. Introduction

World Wide Web is no doubt a major source of information that creates new challenges of information retrieval. This is because the amount of information on the web is increasing exponentially. Web Mining is defined as the application of data mining techniques to extract and analyze useful information from Web data. Based on the kind of data to be mined, Web Mining can be classified into three different categories. These are: Web Content Mining, Web Structure Mining and Web Usage Mining[1].Web Usage Mining, as a technique of Web Mining, is defined as the process of applying data mining techniques to discover usage patterns from data extracted from Web Log files. In Web Usage Mining, data mining techniques are applied to pre-processed Web log data with the purpose of finding interesting and useful patterns. The web log files on the web server are considered to be the major source of data for Web Usage Mining. Web usage mining techniques can be used to automatically extract frequent access patterns from the history of previous user click streams stored in Web log files. The Web Usage Mining process consists of these steps: Data Collection, Pre-processing, Pattern Discovery and Pattern Analysis[2]. Web log pre-processing is used to clean raw log data. The purpose of using data pre-processing is to extract useful data from raw web log and then transform these data into a form necessary for pattern identification[3]. Since the origin web logs data sources are mixed with irrelevant information, data pre-processing acts as an important step to filter and organize only suitable and relevant information before presenting it to any web mining algorithm [4]. In the pre-processing process, a series of processing tasks are applied on web log file including data cleaning, user identification, session identification[5].The rest of the paper is organized as follows: Section 2 describes various works done in this area. Section 3 describes the steps of log data pre-processing aimed to identify data cleaning, user identification and user session.

Finally, Section 4 presents experimental results and draws final conclusions.

2. The Usage Mining on the Web

Web Usage Mining (WUM) is the process of applying data mining techniques to the discovery of usage patterns from data extracted from Web log files. It mines secondary data (web logs) derived from the users' interaction with web pages during certain period of Web sessions. Web usage mining consists of three phases, namely: pre-processing, pattern discovery, and pattern analysis[6 - 10], as shown in figure1.
The goal of web usage mining is to get into the records of servers (log files) that store the transactions performed in the web in order to find patterns revealing the usage of the customers [11,12]. WUM has become an active area of research in the field of data mining due to its vital importance [13].

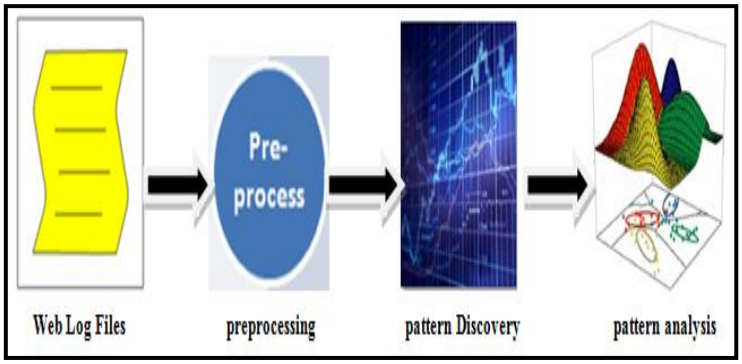


Fig. 1 Phases of Web Usage Mining

2.1 Web Access Patterns

Web access pattern mining is an application of sequence mining on web log data to generate interesting user access behaviors on World Wide Web. Mining of web access patterns generated by the users' interaction with the World Wide Web is thrust area of research.

2.2 The Log File: what is it and how do we store information on it?

2.2.1 The log file's definition

A log file is defined as "a file that lists actions that have occurred" [14]. Such files are generated by servers – a computer or a device on a network that manages network resources and contains a list of all requests made to the server by the network's users.
A Web log file [15] records activity information when a Web user submits a request to a Web Server. The main source of raw data is the web access log which we shall refer to as the log file.

2.2.2 The way we store information on a log file

As it is the rule for every file, information in the log file has to be written in a specific format; that is in a specific sequence and in a certain way that will facilitate the analysis of the file and 'instruct' the computer as to how to read and use[14]. Log files can be located in three places [16,17].

- **Web Servers-** A web server dispenses the web pages as they are requested
- **Proxy Server-** A proxy server is an intermediary computer that acts as a computer hub through which user requests are processed.
- **Web Client-** A Web client is a computer application, such as a web browser, that runs on a user local computer or workstation and connects to a server as necessary.

2.3 Web Server Log File

A web server log file is a log file that automatically creates and maintains the activities performed in it. This file is used to record each and every hit to a web site[18]. It maintains a history of page requests, also helps us in understanding how and when your website pages and application are being accessed by the web browser. These log files contain information such as an IP address of a remote host, content requested, and time of request.

2.4 NCSA Combined Log Format

National Centre for Supercomputing Application (NCSA),stores all common log information with two additional fields. referrer and user agent.
Syntax: Host IP address, Proprietor, Username, date: time, request method, status code, byte size, referrer, and user agent.

3. Data Pre-processing on Web Server Logs

Web usage mining is the application of data mining techniques to usage logs of large data repositories. Usually, the data collected in web log file is incomplete and not suitable for mining directly. Therefore, pre-processing is necessary to convert the data into a suitable form for pattern discovery[19].We begin this phase by data extraction then data cleaning and finally data filtering, because the origin web logs data sources are blended with irrelevant information. Data pre-processing plays an important role in Web usage mining . It uses to filter and organize only appropriate information before using Web mining algorithms on the Web server logs.
The original server logs are cleaned, formatted, and then grouped into meaningful sessions before being utilized by WUM.
This phase contains three sub-steps: Data Cleaning, User Identification, and Session Identification as shown in figure 2.

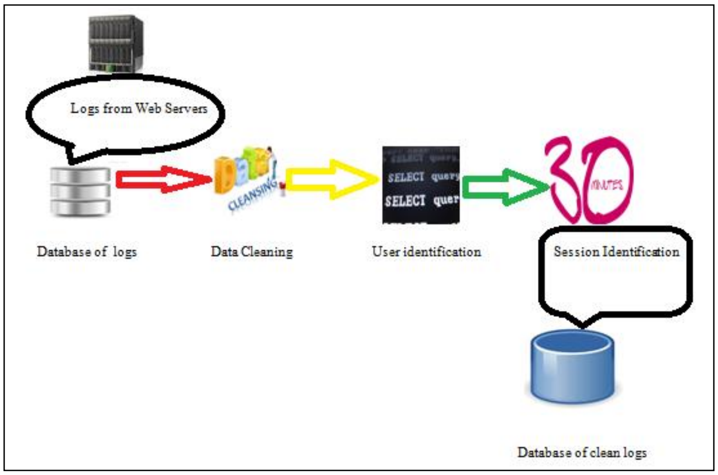


Fig. 2 Pre-processing steps

3.1 Data Cleaning

The data cleaning process removes the data tracked in Web logs that are useless or irrelevant for mining purposes. The request processed by auto search engines, such as Crawler, Spider, and Robot, and requests for graphical page content. Thus the data cleaning step removes the following entries from the original log file[20].

- The entries having suffixes like .jpg, .jpeg, .css, .map etc.,
- Entries having status code failure.
- Remove all record which do not contain method " GET".
- Remove navigation sessions performed by Crawler ,Spider, and Robot.

3.2 User Identification

User identification is the process of identifying each different user accessing Web site. Goal of user identification is to mine every user's access characteristic, and then make user clustering and provide personal service for the users. Each user has unique IP address and each IP address represents one user. But in fact there are three conditions : (l) Some users has unique IP address. (2) Some user has two or more IP addresses. (3) Due to proxy server, some user may share one IP address. Rules for user identification are:

- Different IP addresses refer to different users.
- The same IP with different operating systems or different browsers should be considered as different users.
- While the IP address, operating system and browsers are all the same, new user can be determined whether the requesting page can be reached by accessed pages before, according to the topology of the site.
- Users are uniquely identified by combination of referrer URL and user agent.

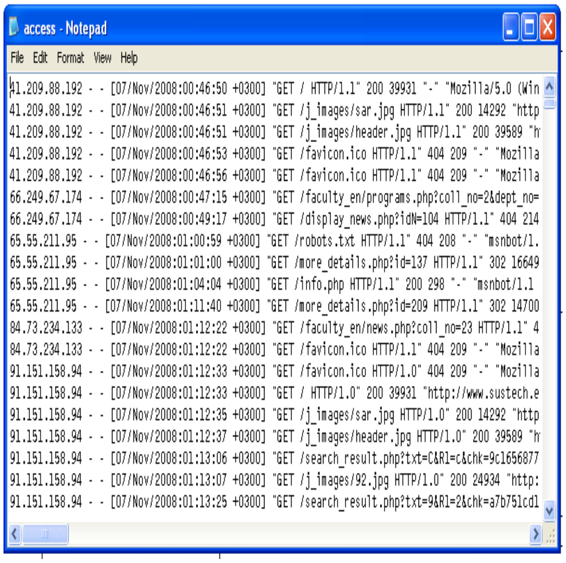
3.3 User session

After identifying users, we need to identify sessions. To do this, we can divide access of the same users into sessions. It is difficult to detect when one session is finished and start another. To detect sessions is common use of time between requests; if two requests are called in of time frame, we can suppose that these requests are in the same session; in another way below of time frame, we can consider two different sessions. A good time frame is 30 minutes.

4. Experimental Results

4.1 Source of the data

As the developed system is to be used to identify trends of visitor website behavior within the SUST web site applications from 7/Nov/2008 through 20/Dec/2009. A portion of the log file used for the experimentation has been shown in the following figure 3.



```
41.209.88.192 - - [07/Nov/2008:00:46:50 +0300] "GET / HTTP/1.1" 200 39931 "-" "Mozilla/5.0 (Win
41.209.88.192 - - [07/Nov/2008:00:46:51 +0300] "GET /j_images/sar.jpg HTTP/1.1" 200 14292 "http
41.209.88.192 - - [07/Nov/2008:00:46:51 +0300] "GET /j_images/header.jpg HTTP/1.1" 200 39589 "h
41.209.88.192 - - [07/Nov/2008:00:46:53 +0300] "GET /favicon.ico HTTP/1.1" 404 209 "-" "Mozilla
41.209.88.192 - - [07/Nov/2008:00:46:56 +0300] "GET /favicon.ico HTTP/1.1" 404 209 "-" "Mozilla
66.249.67.174 - - [07/Nov/2008:00:47:15 +0300] "GET /faculty_en/programs.php?coll_no=2&dept_no=
66.249.67.174 - - [07/Nov/2008:00:49:17 +0300] "GET /display_news.php?idN=104 HTTP/1.1" 404 214
65.55.211.95 - - [07/Nov/2008:01:00:59 +0300] "GET /robots.txt HTTP/1.1" 404 208 "-" "msnbot/1.
65.55.211.95 - - [07/Nov/2008:01:01:00 +0300] "GET /more_details.php?id=137 HTTP/1.1" 302 16649
65.55.211.95 - - [07/Nov/2008:01:04:04 +0300] "GET /info.php HTTP/1.1" 200 298 "-" "msnbot/1.1
65.55.211.95 - - [07/Nov/2008:01:11:40 +0300] "GET /more_details.php?id=209 HTTP/1.1" 302 14700
84.73.234.133 - - [07/Nov/2008:01:12:22 +0300] "GET /faculty_en/news.php?coll_no=23 HTTP/1.1" 4
84.73.234.133 - - [07/Nov/2008:01:12:22 +0300] "GET /favicon.ico HTTP/1.1" 404 209 "-" "Mozilla
91.151.158.94 - - [07/Nov/2008:01:12:33 +0300] "GET /favicon.ico HTTP/1.0" 404 209 "-" "Mozilla
91.151.158.94 - - [07/Nov/2008:01:12:33 +0300] "GET / HTTP/1.0" 200 39931 "http://www.sustech.e
91.151.158.94 - - [07/Nov/2008:01:12:35 +0300] "GET /j_images/sar.jpg HTTP/1.0" 200 14292 "http
91.151.158.94 - - [07/Nov/2008:01:12:37 +0300] "GET /j_images/header.jpg HTTP/1.0" 200 39589 "h
91.151.158.94 - - [07/Nov/2008:01:13:06 +0300] "GET /search_result.php?txt=C&R1=c&chk=9c1656877
91.151.158.94 - - [07/Nov/2008:01:13:07 +0300] "GET /j_images/92.jpg HTTP/1.0" 200 24934 "http:
91.151.158.94 - - [07/Nov/2008:01:13:25 +0300] "GET /search_result.php?txt=9&R1=2&chk=a7b751cd1
```

Fig. 3 Extract of SUST log file

Descriptions of the access that were utilized to generate the data sets are provided below, each row of the log contains the information is shown in table 1.

Table. 1 Description of log access used to generate data sets

Access name	**Description**
IP Address	Remote IP address
Proprietor	The name of the owner making an http request
User name	Username and password if the server requires user authentication
Date / Time	date/time of the transaction
Method	Modes of request
URL	URL requested by the client
Protocol	HTTP protocol
Statues code	HTTP return code
Byte Size	Size in bytes of the response sent to the client
Referred	The site from which the visitor came
Agent	User agent

log proprietor- The name of the owner making an http request is recorded through this field. They do not expose this information for security purpose. When they are not exposed they are denoted by (-).
Username- This field records the name of the user when it gets a http request. They do not expose this information for security purpose. When they are not exposed they are denoted by (-).
Figure 4, shows a sample of a single entry log file a common transfer log extract from SUST log file.

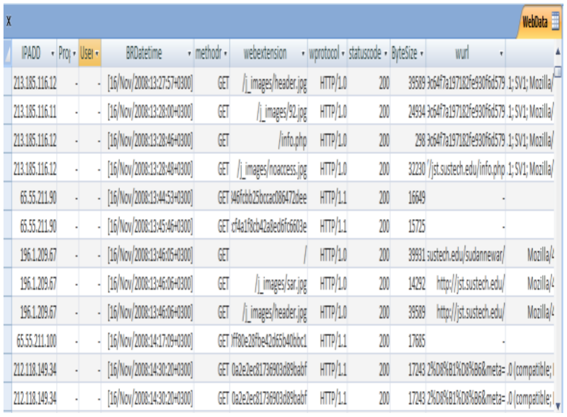
```
41.209.88.192 - - [07/Nov/2008:00:46:51
+0300] "GET /j_images/sar.jpg HTTP/1.1" 200
14292 "http://jst.sustech.edu/" "Mozilla/5.0
(Windows; U; Windows NT 6.0; en-US;
rv:1.9.0.3) Gecko/2008092417 Firefox/3.0.3"
```

Fig. 4 Single entry of log file from SUST

Here, 41.209.88.192 is the IP address of the client, 07/Nov/2008:00:46:51 is the date/time of transaction; GET is the method of transaction, j_images/sar.jpg is URL requested by client, HTTP/1.1 is the HTTP protocol,200 is HTTP return code (200 means OK), 14292 is the size in bytes of the response sent to the client, http://jst.sustech.edu is the URL referring to the request one, “Mozilla/5.0 (Windows; U; Windows NT 6.0; en-US; rv:1.9.0.3) Gecko/2008092417 Firefox/3.0.3" is the user agent.

4.2 Data Field Extraction and Transfer Server Logs to database

A server log file consists of various data fields that should be separated before applying any cleaning procedure. The process of separating out different data fields from single server log entry is identified as data field extraction. A server uses different characters such as a comma or a space character which works as separators.
To analyze the log file data, we need first to deal with text file using regular expression so as to separate each line in the text file into different fields and then we need database object for loading these data in a database table. The whole log file can be read in one variable and then we can move this variable line by line using a loop statement.
After reading the log files, several attributes are considered important for the analysis. The read logs records will be stored in a database. Figure 5 shows the database to store the data.



IPADD	Pro	Use	BRDatetime	methodr	webextension	wprotocol	statuscode	ByteSize	wurl	
213.185.116.12	-	-	[16/Nov/2008:13:27:57+0300]	GET	/j_images/header.jpg	HTTP/1.0	200	39589	3c64f7a197182fe930f6d579	.1; SV1; Mozilla/
213.185.116.11	-	-	[16/Nov/2008:13:28:00+0300]	GET	/j_images/92.jpg	HTTP/1.0	200	24934	3c64f7a197182fe930f6d579	.1; SV1; Mozilla/
213.185.116.12	-	-	[16/Nov/2008:13:28:46+0300]	GET	/info.php	HTTP/1.0	200	298	3c64f7a197182fe930f6d579	.1; SV1; Mozilla/
213.185.116.12	-	-	[16/Nov/2008:13:28:48+0300]	GET	/j_images/noaccess.jpg	HTTP/1.0	200	32230	'/jst.sustech.edu/info.php	.1; SV1; Mozilla/
65.55.211.90	-	-	[16/Nov/2008:13:44:53+0300]	GET	946fcbb25bccad086472dee	HTTP/1.1	200	16649	-	
65.55.211.90	-	-	[16/Nov/2008:13:45:46+0300]	GET	cf4a1f8cb42a8ed6fc6603e	HTTP/1.1	200	15725	-	
196.1.209.67	-	-	[16/Nov/2008:13:46:05+0300]	GET	/	HTTP/1.0	200	39931	sustech.edu/sudannewar/	Mozilla/4
196.1.209.67	-	-	[16/Nov/2008:13:46:06+0300]	GET	/j_images/sar.jpg	HTTP/1.0	200	14292	http://jst.sustech.edu/	Mozilla/4
196.1.209.67	-	-	[16/Nov/2008:13:46:06+0300]	GET	/j_images/header.jpg	HTTP/1.0	200	39589	http://jst.sustech.edu/	Mozilla/4
65.55.211.100	-	-	[16/Nov/2008:14:17:09+0300]	GET	)ff80e28fbe42d65b40bbc1	HTTP/1.1	200	17685	-	
212.118.149.34	-	-	[16/Nov/2008:14:30:20+0300]	GET	0a2e2ec81736903d89babf	HTTP/1.1	200	17243	2%D8%B1%D8%B6&meta=	.0 (compatible;
212.118.149.34	-	-	[16/Nov/2008:14:30:20+0300]	GET	0a2e2ec81736903d89babf	HTTP/1.1	200	17243	2%D8%B1%D8%B6&meta=	.0 (compatible;

Fig. 5 Table after data is transferred to database

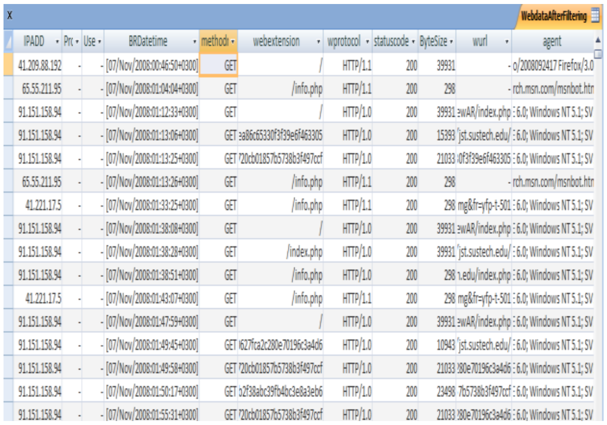
Figure 5 shows the server log data after transferring to database and note that all attributes are shown in this figure due to the space restrictions. Several attributes are interesting fields are included in the database.

4.3 Data Cleaning

After data cleaning only 122122 entries are left in the log. The results are shown in figure 6.The VB.Net is use to implements this function is written in steps as:

```
1.  Define variables (method, status code, agent
    and web extension) As string

2.  Check method:
       If method = "GET" Then
          method = 1
       Else
          method = 0
       End If

3.  Check status code
       If status code = "200" Then
          status code = 1
       Else
          status code= 0
      End If

4.  Check agent
       If agent contain the Spider Or Robot Or Crawler Then
          agent = 0
       Else
          agent = 1
       End If

5.  check web extension
       If web extension .jpgi Or .jpegi Or .jsi Or .cssi Or .gifi Then
          web extension= 0
       Else
          web extension = 1
       End If

6.  Add data
       If method = 1 And web extension = 1 And status code = 1 And agent = 1
       Then
      "INSERT INTO WebdataAfterFiltering(all fields)
       End If
    Next i

7.  Close db
```

Figure 6 shows the server log data after data cleaning.



IPADD	Prc	Use	BRDatetime	methodı	webextension	wprotocol	statuscode	ByteSize	wurl	agent
41.209.88.192	-	-	[07/Nov/2008:00:46:50+0300]	GET	/	HTTP/1.1	200	39931	-	o/2008092417 Firefox/3.0
65.55.211.95	-	-	[07/Nov/2008:01:04:04+0300]	GET	/info.php	HTTP/1.1	200	298	-	rch.msn.com/msnbot.htn
91.151.158.94	-	-	[07/Nov/2008:01:12:33+0300]	GET	/	HTTP/1.0	200	39931	ewAR/index.php	:6.0; Windows NT 5.1; SV
91.151.158.94	-	-	[07/Nov/2008:01:13:06+0300]	GET	ea86c65330f3f39e6f463305	HTTP/1.0	200	15393	'jst.sustech.edu/	:6.0; Windows NT 5.1; SV
91.151.158.94	-	-	[07/Nov/2008:01:13:25+0300]	GET	'20cb01857b5738b3f497ccf	HTTP/1.0	200	21033	0f3f39e6f463305	:6.0; Windows NT 5.1; SV
65.55.211.95	-	-	[07/Nov/2008:01:13:26+0300]	GET	/info.php	HTTP/1.1	200	298	-	rch.msn.com/msnbot.htn
41.221.17.5	-	-	[07/Nov/2008:01:33:25+0300]	GET	/info.php	HTTP/1.1	200	298	mg&fr=yfp-t-501	:6.0; Windows NT 5.1; SV
91.151.158.94	-	-	[07/Nov/2008:01:38:08+0300]	GET	/	HTTP/1.0	200	39931	ewAR/index.php	:6.0; Windows NT 5.1; SV
91.151.158.94	-	-	[07/Nov/2008:01:38:28+0300]	GET	/index.php	HTTP/1.0	200	39931	'jst.sustech.edu/	:6.0; Windows NT 5.1; SV
91.151.158.94	-	-	[07/Nov/2008:01:38:51+0300]	GET	/info.php	HTTP/1.0	200	298	1.edu/index.php	:6.0; Windows NT 5.1; SV
41.221.17.5	-	-	[07/Nov/2008:01:43:07+0300]	GET	/info.php	HTTP/1.1	200	298	mg&fr=yfp-t-501	:6.0; Windows NT 5.1; SV
91.151.158.94	-	-	[07/Nov/2008:01:47:59+0300]	GET	/	HTTP/1.0	200	39931	ewAR/index.php	:6.0; Windows NT 5.1; SV
91.151.158.94	-	-	[07/Nov/2008:01:49:45+0300]	GET	i627fca2c280e70196c3a4d6	HTTP/1.0	200	10943	'jst.sustech.edu/	:6.0; Windows NT 5.1; SV
91.151.158.94	-	-	[07/Nov/2008:01:49:58+0300]	GET	'20cb01857b5738b3f497ccf	HTTP/1.0	200	21033	'80e70196c3a4d6	:6.0; Windows NT 5.1; SV
91.151.158.94	-	-	[07/Nov/2008:01:50:17+0300]	GET	b2f38abc39fb4bc3e8a3eb6	HTTP/1.0	200	23498	7b5738b3f497ccf	:6.0; Windows NT 5.1; SV
91.151.158.94	-	-	[07/Nov/2008:01:55:31+0300]	GET	'20cb01857b5738b3f497ccf	HTTP/1.0	200	21033	'80e70196c3a4d6	:6.0; Windows NT 5.1; SV

Fig 6. Result after data cleaning

Table 3 shows the comparison between size and number of records before and after data cleaning. Figures 7 and 8 illustrate the change in the number of records and log file size, respectively.

Table. 3 Size and number of records before and after cleaning

	Size(MB)	Number of records
before	567	291642
After	143	122122
Reduced	424	169520
Percentage in Reduction	74.77%	58.13%

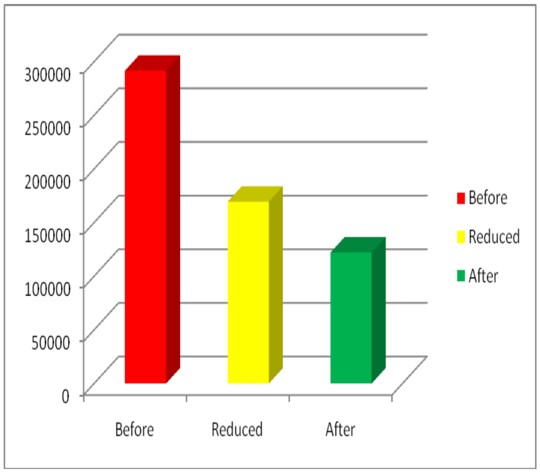


Fig 7. Bar Chart Showing the Change in Number of Records

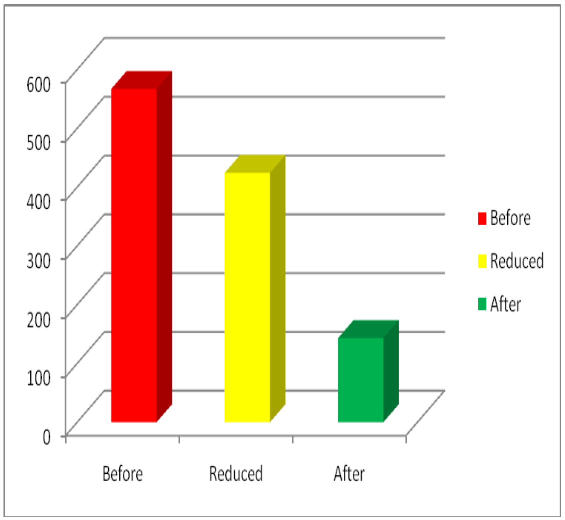


Fig. 8 Bar Chart Showing Change in log Size(MB)

4.4 User identification

Compute the unique user by combination of Referred and Agent. Table 4 shows the number of records, Unique IP address and Unique users.

a- Distinct IP address refer to different users.

b- Combine Referred and Agent.

c- The same IP with different combined felid should be considered as different users.

Table. 4 Number of record, Unique IP address and Unique users

Total No. of records	122122
Total No. of Unique IP address	11030
Total No. of Unique users	23397

4.5 User session

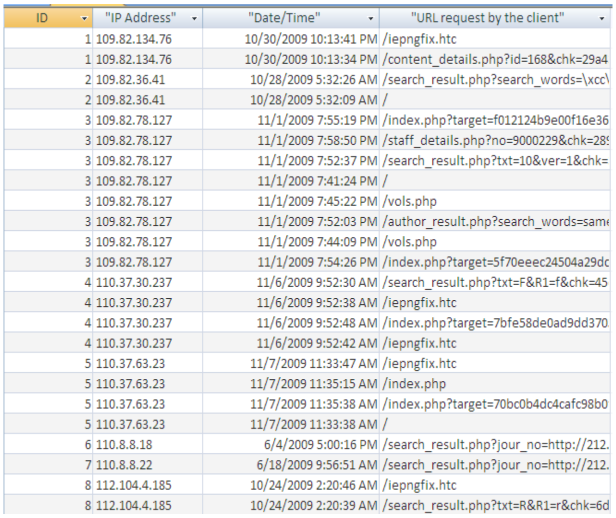
After we specify the number of unique users in the previous step, we need to get the users sessions. To achieve this we can divide the access of the same users into sessions. The time spent within time limit of 30 minutes for same user will be consider a user session. The following is the session algorithm. Rules for session identification are:

- Different IP addresses refer to different session.
- The same user with time exceeds a certain limit (30 mintes) should be considered as different session. The algorithm below applied for this purpose.

Algorithm session

```
Begin
Start session=Current time of current session
Session =1
While not eof (LogFile) Do
LogRecord=Read (LogFile)
In the same IP Address
If (the time of current Record –start session) <= 30 min
Then
We are in same the session
Move to next record
Else
increment session = session +1
Start time = time of current Record
Move  to next record
End If
End While
End
```

A fraction of log files with user sessions identification is shown in figure 8. There were a total of 23200 unique visiting IP addresses, 8861 unique pages and 13869 sessions. Figure 9 shows the session length distribution for the SUST dataset after the session is identified. The vertical axis stands for the percentage of occurrence of the number of session length. The horizontal axis is marked with the length of session.



ID	"IP Address"	"Date/Time"	"URL request by the client"
1	109.82.134.76	10/30/2009 10:13:41 PM	/iepngfix.htc
1	109.82.134.76	10/30/2009 10:13:34 PM	/content_details.php?id=168&chk=29a4
2	109.82.36.41	10/28/2009 5:32:26 AM	/search_result.php?search_words=\xcc\
2	109.82.36.41	10/28/2009 5:32:09 AM	/
3	109.82.78.127	11/1/2009 7:55:19 PM	/index.php?target=f012124b9e00f16e36
3	109.82.78.127	11/1/2009 7:58:50 PM	/staff_details.php?no=9000229&chk=28
3	109.82.78.127	11/1/2009 7:52:37 PM	/search_result.php?txt=10&ver=1&chk=
3	109.82.78.127	11/1/2009 7:41:24 PM	/
3	109.82.78.127	11/1/2009 7:45:22 PM	/vols.php
3	109.82.78.127	11/1/2009 7:52:03 PM	/author_result.php?search_words=same
3	109.82.78.127	11/1/2009 7:44:09 PM	/vols.php
3	109.82.78.127	11/1/2009 7:54:26 PM	/index.php?target=5f70eeec24504a29dc
4	110.37.30.237	11/6/2009 9:52:30 AM	/search_result.php?txt=F&R1=f&chk=45
4	110.37.30.237	11/6/2009 9:52:38 AM	/iepngfix.htc
4	110.37.30.237	11/6/2009 9:52:48 AM	/index.php?target=7bfe58de0ad9dd370
4	110.37.30.237	11/6/2009 9:52:42 AM	/iepngfix.htc
5	110.37.63.23	11/7/2009 11:33:47 AM	/iepngfix.htc
5	110.37.63.23	11/7/2009 11:35:15 AM	/index.php
5	110.37.63.23	11/7/2009 11:35:38 AM	/index.php?target=70bc0b4dc4cafc98b0
5	110.37.63.23	11/7/2009 11:33:38 AM	/
6	110.8.8.18	6/4/2009 5:00:16 PM	/search_result.php?jour_no=http://212.
7	110.8.8.22	6/18/2009 9:56:51 AM	/search_result.php?jour_no=http://212.
8	112.104.4.185	10/24/2009 2:20:46 AM	/iepngfix.htc
8	112.104.4.185	10/24/2009 2:20:39 AM	/search_result.php?txt=R&R1=r&chk=6d

Fig. 9 A fragment from users session result.

The result in figure 10 shows that a significant number of sessions only consist of one or two request. and there are not too many web user sessions which extend over 5 visits.

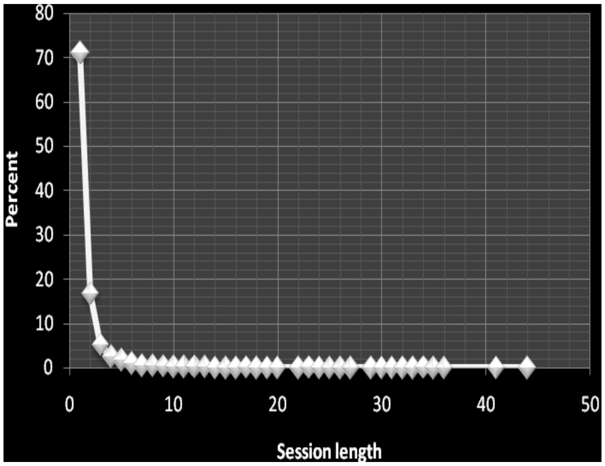


Fig. 10 The session length distribution of data set

5. Conclusions

Data pr-processing is one of the important and prerequisite phases in WUM. Web log files are the best source to predict user's behavior. Along with the useful information, the raw log files also contain entries for unnecessary details like image access, failed entries etc. Many interesting patterns are available in the Web log data. However, it is very complicated to extract the interesting patterns without the pre-processing phase. This paper presents a brief introduction to WUM, apart from the data mining technologies and also the implementation of the pre-processing of Web log files in SUST Web server. This study focuses on methods that can be used for the tasks of data cleaning, user identification and session identification from Web log files. The results obtained after preprocessing were satisfactory and contained valuable information about the log files. The results showed a reduction in the number of records and in the log file size and hence increases the quality of the available data. This paper gives a detailed look about Web mining, Web using mining , Web server log file and its format and data pre-processing .The pre-processed data is then made available for further pattern discovery and pattern analysis.

Acknowledgments

I thank Dr. Murtada Elfaki for many useful and constructive comments and suggestions, which helped me to significantly improve my work.

References

[1] A. Sharma, A. Kumar, and Dr PC Gupta, "Exploration of efficient methodologies for the Improvement in web mining techniques: A survey," International Journal of Research in IT & Management 1.3 (2011): 85-95.

[2] A. Al-qwaqenah and M. Al-kabi, “Discovering the Web Usage in three Jordanian Universities,”The International Conference on Information and Communication Systems, 2011,pp. 1–8.

[3] R. Gupta and P. Gupta, “Application specific web log pre-processing,” Int.J. Computer Technology & Applications, vol. 3, no. 1, 2012, pp. 160–162.

[4] M. Helmy, A. Wahab, M. Norzali, H. Mohd, H. F. Hanafi, M. Farhan, and A. Background, “Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm,” Proceedings Of World Academy of Science, Engineering and Technology, vol. 36, no. December, 2008, pp. 970–977.

[5] P. Nithya and P. Sumathi, “An Effective Web Usage Analysis using Fuzzy Clustering,” ARPN Journal of Science and Technology, vol. 3, no. 7, 2013,pp. 693–698.

[6] K. Etminani, “Web Usage Mining: users’ navigational patterns extraction from web logs using Ant-based Clustering Method,”IFSA-EUSFLAT,2009,pp. 396–401.

[7] R. Gupta and P. Gupta, “Application specific web log pre-processing,” Int. J.Computer Technology & Applications, vol. 3, no. 1, pp. 2012,160–162.

[8] R. Gupta and P. Gupta, “Fast Processing of Web Usage Mining with Customized Web Log Pre-processing and modified Frequent Pattern Tree,” International Journal of Science and Communication Networks, vol. 1, no. 3, pp. 2011,277–279.

[9] D. S. Singh, Arun, Avinav Pathak, “Web Usage Mining : Discovery Of Mined Data Patterns and their Applications,” International Journal of Computer Science and Management Research, vol. 2, no. 5, 2013,pp. 2423–2429.

[10] M. S. Kamat, J. W. Bakal, and M. Nashipudi, “Improved Data Preparation Technique in Web Usage Mining,” International Journal of Computer Networks and Communications Security, vol. 1, no. 7, 2013, pp. 284–291.

[11] S. K. Pani , L.Panigrahy, V.H.Sankar, Bikram Keshari Ratha, A.K.Mandal, S.K.Padhi,” Web Usage Mining: A Survey on Pattern Extraction from Web Logs”, International Journal of Instrumentation, Control & Automation , Volume 1, Issue 1, 2011,pp.15-23.

[12] Sawan Bhawsar, Kshitij Pathak, Sourabh Mariya, Sunil Parihar," Extraction of Business Rules from Web logs to improve Web Usage Mining", Vol 2, Issue 8, Aug 2012,pp.333-340.

[13] P. Nithya and P. Sumathi, “A Survey on Web Usage Mining: Theory and Applications,” International Journal, vol. 3, no. August 2012, pp. 1625–1629.

[14] G. K. Lekeas, “Data mining the web: the case of City University’s Log Files,” 2000.

[15] Stava, Jaideep, et al. “Web usage mining: Discovery and Applications of usage patterns from web data,”, ACM SIGKDD Explorations Newsletter 1.2 (2000): 12-23.

[16] K. Suneetha and R. Krishnamoorthi, “Identifying user behavior by analyzing web server access log file,” IJCSNS International Journal of Computer Science and Network Security”, vol. 9, no. 4, 2009, pp. 327–332.

[17] S. G. Langhnoja, M. P. Barot, and D. B. Mehta, “Web Usage Mining Using Association Rule Mining on Clustered Data for,” International Journal of Data Mining Techniques and Applications, vol. 02, no. 01, 2013, pp. 141–150.

[18] O. CU and P. Bhargavi, “Analysis of Web Server Log by Web usage Mining for Extracting users Patterns,” http://www. tjprc.org/view_archives. Php, vol. 3, no. 2, pp. 123–136, 2013.

[19] K. B. Patel, “Process of Web Usage Mining to find Interesting Patterns from Web Usage Data,” International Journal of Computer Applications & Technology, vol. 3, no. 1, 2012, pp. 144–148.

[20] P. Nithya and P. Sumathi, “An Effective Web Usage Analysis using Fuzzy Clustering,” ARPN Journal of Science and Technology, vol. 3, no. 7, 2013,pp. 693–698.