International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-9 Issue-1, May 2020

The Role of Data Preprocessing System on Web Log Files for Mining Students Access Logs

A. Dhana Praveena, V. Selvi

Abstract: ***Nowadays, WWW has grown into significant and vast data storage. Every one of clients' exercises will be put away in log record. The log file shows the eagerness on the website. With an abundant use of web, the log file size is developing hurriedly. Web mining is a utilization of information digging innovations for immense information storehouses. It is the procedure of uncover data from web information. Before applying web mining procedures, the information in the web log must be pre-processed, consolidated and changed. It is essential for the web excavators to use smart apparatuses so as to discover, concentrate, channel and assess the ideal data. The information preprocessing stage is the most significant stage during the time spent web mining and is basic and complex in fruitful extraction of helpful information. The web logs are circulated in nature also they are non-versatile and unfeasible. Subsequently we require a broad learning calculation so as to get the ideal data.***

Keywords: Cleaning, Pre-processing, Session identification, User identification, Web log files.

I. INTRODUCTION

In data mining system, Web mining is an application utilized for huge web data repositories. It is applied to determine unknown patterns also connections surrounded by the web data. Weblog mining is recognized to find the real world problems such as mostly clicked pages, continual accesses, interesting user plus user performance emerging patterns so it will aid the creator to progress more themes. It can be used in a variety of claims such as structure enhancement, website alteration, commercial intelligence and so on. This information can be used to take revamp decisions. The common method of web mining contains,

- *Resource collection:* Process of extracting the task relevant data,
- *Information pre processing:* Process of cleaning, Integrating and Transforming of the result of resource collection,
- *Pattern discovery:* Process of uncovered general patterns in the pre process data
- *Pattern analysis:* Process of validating the discovered patterns [1].

Meanwhile web extracting valuable patterns from perusing conduct of clients so as to comprehend their inclinations is turning into a basic reality in the advancement of versatile sites. Web client's exercises can be caught into an extraordinary document called web log file. Web logs are kept up in the web servers as plain content documents which contains the insights regarding client name, IP address, date, time, number of bytes moved, get to ask for and referrer log. The web logs are rundown of page mentions by the clients or snap stream information which contains conflicting and inadequate data. In this way, it is hard to utilize the web logs legitimately for design mining calculations to remove the highlights [2]. Preprocessing strategies are important to make them predictable and finish. It comprises of information cleaning, client distinguishing proof, meeting recognizable proof and information change. Information preprocessing is utilized to clean the superfluous information from log record so it very well may be given to the example disclosure to recognize the client design. There are different kinds of a log:

- *Server log:* When an internet user request a particular page on web, an entry is logged into a special file called server log file. This file is not accessible by general internet user, only administrative person or server owners can access these files.
- *Proxy server log:* A Proxy server is a server which acts as an agent between user's requests to other web servers. They are generally used for caching services to improve navigation speed, administrative control and security.
- *Client/Browser log:* Web log data can also be collected from client machine by integrating java applets to the website, writing java scripts or even modified browsers. Client side logs are useful to tackle problems related with server logs like web page caching, session reconstruction [3].

These log documents are damaged by web use mining to investigate plus discover helpful patterns. Precise distinguishing proof of client perusing designs is especially significant in Web personalization not exclusively to aid the webpage's proprietor in refining its quality yet additionally to adjust the structure of the website.

Manuscript received on April 02, 2020.
Revised Manuscript received on April 20, 2020.
Manuscript published on May 30, 2020.
* Correspondence Author

A. Dhana Praveena*, Research Scholar, Department of Computer Science, Mother Teresa Women's University, Kodaikanal, Tamilnadu, India. Email: dhanapraveena@yahoo.co.in

Dr. V. Selvi, Assistant Professor, Department of Computer Science, Mother Teresa Women's university, Kodaikanal, Tamilnadu, India. Email:selvigiri.s@gmail.com

II. SYSTEM ARCHITECTURE

For the most part the web get to logs are dispersed and quickly developing in environment. It is important for web miners to abuse astute capacities so as to discover, extract, filter and assess the perfect data. Earlier utilizing 'web mining' methods to 'web usage data', the 'web usage store assortment' must be cleansed, coordinated and changed.

a. Need For Preprocessing

A click stream is an arrangement of pages visited by a client through site in a specific timeframe. Each hit against the server relating to an http demand produces a section in server get to log records. Each log section encompasses the date and time of solicitation, the IP address of the customer, the substance mentioned, and status of the demand, the strategy utilized and a lot more items. Each demand is logged discretely, so at the phase of preprocessing these demands must be aggregate so as to structure the information with the goal that it gives some data [4].

b. Description of Data Set

Any educational institution (universities/ colleges) management battles hard in satisfying the expectations of the students conceded. In the midst of the conceded students, some may be interested in jobs; some might be sharp towards look into and different towards advanced education or business enterprise. So as to achieve the objectives, the people pulling the strings of the universities need to comprehend the attitudes of the understudies and shape them subsequently. So as to perceive the student's fantasies, aside from different techniques, for example, addressing, the route conduct of the understudies likewise assumes a crucial job.
In our proposed work the conduct and surfing qualities of students are considered. This work clarifies the way toward recognizing the understudy standard of conduct by following the web perusing conduct of the student covered up in the log records of college's web server. The crude information is preprocessed, to create sorted out data. The log document positions are recorded beneath,

- Common Web Log Format (CLF)
- W3C Extended log file format
- Microsoft IIS(Internet Information Server) web log file format
- NCSA (National Center for Supercomputing Application) Common web log file format
- Apache Log File Format [4].

A 'web log' typically shields accesses with esteem to

Web log activities

Host IP Address

User Authentication

Date plus Time of visit

HTTP Request

Referrer Field

User Agent Field

W3C Extended log format

c. Proposed system:

- Common Web Log Format (CLF)
- W3C Extended log file format
- Microsoft IIS(Internet Information Server) web log file format
- NCSA (National Center for Supercomputing Application) Common web log file format
- Apache Log File Format [4].

The smart method gets the fresh web log as input as well as rejects the web search engine admissions routinely with small period. It produces preferred web log contains of only human user accesses. The web log preprocessing construction shown in Figure 1.

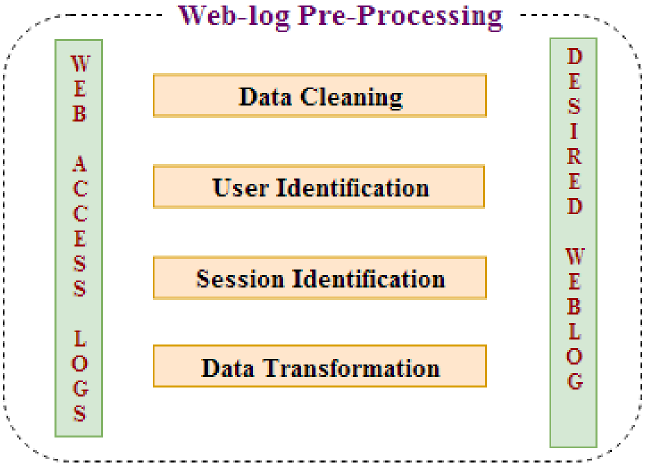


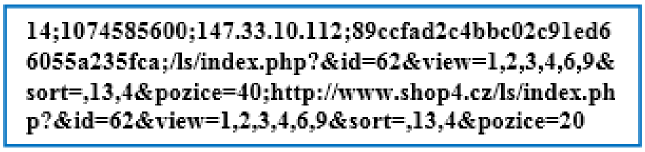
Fig1. Steps in Web log Pre-Processing

In this paper the following steps are consisted for experiment.

i. **Data (Web access log file) collection:** It is the initial level of pre-processing. The user interaction features with the website can be got from Web Servers', 'Proxy Servers' and 'Client Side'. The web logs has 'W3C', 'IIS', NCSA /even the custom format when I is obtained [5]. In this research, 'Web Log Format' has been reserved. Log 'L' can be considered using

$$L=\{(u_1,p_1,t_1,a_1),(u_2,p_2,t_3,a_2)............(u_n,p_n,t_n,s_n,a_n)\}$$

Where 'u_n' refers to the 'n^{th}' user retrieving from the page ' p_n' at timestamp 't_n' with the support of user agent 'a_n'.



```
14;1074585600;147.33.10.112;89ccfad2c4bbc02c91ed6
6055a235fca;/ls/index.php?&id=62&view=1,2,3,4,6,9&
sort=,13,4&pozice=40;http://www.shop4.cz/ls/index.ph
p?&id=62&view=1,2,3,4,6,9&sort=,13,4&pozice=20
```

Fig2. Example Web Log File

The data details are given below:

- *IP address:* It is used to identify the web user visiting the web site.
- *Genuine web User (GU):* It covers the Username and password of the user visiting web site.
- *Date and time of the visit:* It shown user has visited the website timestamp.
- *HTTP Request:* It denotes dataset from the Request Method (GET, POST, HEAD, etc.), the Requested Resource (a HTML page, an image, CSS, CGI, or a script, etc.) and the Protocol Version (HTTP protocol being used along with version number) [5].

Algorithm for data collection:

```
Input: Web Logs (Web Server and Cookies)
Output: Merged Log File
Step 1: Read the log files
Step 2: Distinct the fields either ',' or 'space'
Step 3: Hold the fields in the 'database'
Step 4: Category the fields in rising order by
'time' and 'IP addresses'.
Step 5: If 'IP addresses, 'url' and 'time' are
general for more than one file then remove the
identical record
Step 6: Hold the merged log file
```

ii. Data cleaning:

In preprocessing system, the unrelated data from the 'web log files' are eliminated. It supports to diminish the size of data by eliminating undesirable input log data as well as develop the excellence of (web log files) data. Then further process can only be utilized for filtered data. For instance the records referring images, graphics or video etc. are removed. And also the records with failed HTTP status codes, without proper website link are eliminated [7].

In the proposed Data Cleaning Algorithm, it gives the Log Table that consisting relevant set of records with the below entries as

User IP address	*Date*
Time	*URL details*

By straining out the irrelevant entries, the size of data (web log files) decreases to more than 60% of its initial size. The algorithm for data cleaning is given below.

Algorithm for data cleaning

```
Input: Web Logs (Merged Log File)
Output: Filter Log Table
Step 1: Read the web logs record
Step 2: If ('suffix.url' equals image plus 'multimedia'
file extensions) then
Step 3: Check the previous requests from the same IP
address
Step 4: If ('suffix. Previous' demands equals image
plus 'multimedia' file) then
    the request is explicit request
    add the records to Filter Log Table
    else
    the request is implicit request
    discard the requests
else
    If ('position code' not equivalent to 'miscarriage')
and
    ('user agent' not equivalent to 'crawler', 'spider',
'robot') and (method equals 'GET') then
    Add the records to 'Filter Log Table'
    Step 4: Echo 'Step 2, 3 and 4' till 'end of log File'
```

User Identification: After the data cleaning, the next important steps is identification of users from web log files. Unique IP address can be allocated by the server ('cookies') to identify the operator. This step is essential to classify 'User accesses' of webpages. Every operator will be recognized through a distinctive 'IP address' also if various operators may have similar IP address then the various browser plus various OS (operating system) signify various user [8].

The user ID is to mine all handler's admittance representative, then create 'user clustering' also deliver 'personal service' for the operators. All user has single 'IP address' and every 'IP address' signifies individual operator. However, in detail there are 3 states such as:

- Certain users have single IP address.
- Certain user has more than 2 'IP addresses'.
- Owing to 'proxy server', certain user can share particular 'IP address'.

The user identification algorithm is described in this section. The outcome of this algorithm gives information about,

Total number of individual users

Users IP addresses

User agents

browsers used

Algorithm for user identification

```
Input: Filter Log Table
Output: Log file with distinguished users
Step 1: If 'IP address' is matchless then original
user
Step 2: If 'IP address' is similar then
    user name is not matchless,
    agent log, operating system and browser are
different then distinguish users.
Step 3: Make Web site topology to validate
'access path' also 'identify users'.
Step 4: Echo step 1, 2 and 3 up to 'end of Filter
Log Table'
```

Retrieval Number: F8354038620/2020©BEIESP
DOI:10.35940/ijrte.F8354.059120
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Session Identification: It is the main part of web log 'pre-processing' system for create and visit all specific 'user session'. The objective of this stage is to set page accesses and activities of whole operator into single session. Some of the general techniques utilized to recognize session is break contrivance. The identification method of the session timing is computed by capturing the time transformation among sites also the whole amount of 'clicks' on a specific website certain in a log file [8].

- Group of activities executed by a user from the moment he entered the website to the moment he left it.
- A set of user clicks usually referred to as a click stream, across Web servers is defined as a user session.
- Successions of web pages user browse in a single access.

Traditional session identification technique is based on an even and static timeout. While the interlude between two consecutive requests surpasses the timeout, new session is resolute. Two General Approaches

- Time-oriented heuristic methods
- Navigation-oriented heuristic methods

The beginning time as well as close time for every client process was distinguished also changed over them into particular 20 minutes defaulting periods. The proposed 'User Session Identification' method is described underneath.

Algorithm for session identification

Input: Log Table after User Identification

Output: User interested page identification (based on session time)

```
Step1: Load Logfile using StreamReader
        SELECT cs (Cookie), c-ip,time,
    cs-uri-stem, cs (User-Agent)from
    <user_identified_file> GROUP
        BY cs(Cookie), c-ip, time, cs-uri-stem,
    cs(User-Agent) ORDER BY c-ip, time.
Step2: User path identification purpose
    initializes the variables
Step3: Identify the user viewing time and
    memory
Step4: Read until end of file is encountered.
Step5: Spread this 'data' in 'new file'.
Step6: illustrates the outcomes.
```

Path completion and Data Transformation: There are odds of missing pages subsequent to building exchanges because of intermediary (proxy) servers also storing issues in 'web server logs'. In that state it gets compulsory for recognizing the client's entrance way, and including the missing ways. On account of nearby buffer existence, certain mentioned pages are not recorded in get to log. So as to examine the information in a legitimate manner these 'missing pages' should be affixed. Different approaches for reaching in the omitted pages have been recommend.

In that fruition method, it recognizes distinctive 'user session's' periods by including significant gets to which are not logged in the entrance log. At last, in the 'data pre-processing' task, the grouping of recognized pattern might be put away in the applicable data structures.

Algorithm for Path completion

Input: Output file of 'Session Identification' algorithm.

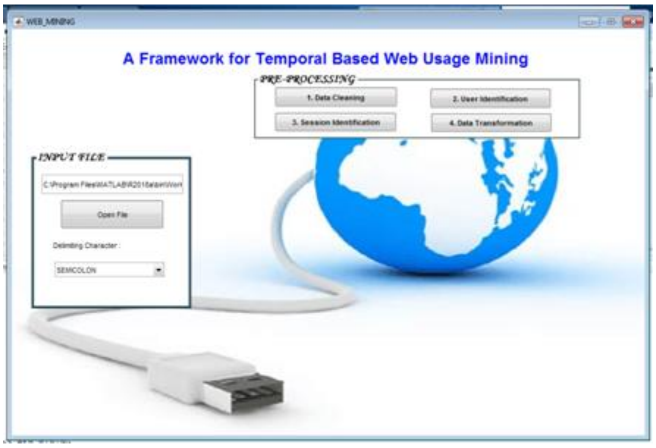
Output: log database without missing value

```
Step1: SELECT cs(Cookie), c-ip, time,
cs-uri-stem,
        cs(User-Agent) from
<sessions_extracted_file>
        GROUP BY cs(Cookie), c-ip, time,
cs-uri-stem,
        cs (User-Agent) ORDER BY cs(Cookie),
c-ip, time.
Step2: Read until end of file is encountered.
Step3: Export this data in new file.
Step4: Show results.
```

After 'path completion' process, data 'renovation procedure' will be utilized for formatting 'web log database'. 'Web log files' will be consisting more amount of records every of that signifies many type of statistics. For example it will convert problematic to contract with various types of data, so the 'transformation step' can be utilized on 'web log files' to convert the data into the layout that is simple as well as appropriate to process in additional stages [9].

III. RESULT AND DISCUSSION

We have utilized student 'web log dataset' in this research. The web log records were gathered from college web server also program machine during '01st January 2019' to'31st June 2019'. It is accessible in stxavierstn.edu.in. It is in 'Web Log Format'. It comprises of a half year of log information. We have utilized MATLAB for our investigation. After the information was recorded it was sent for handling in the instrument planned. The yields created in the wake of tapping on each tab appeared in the proposed work have been appeared. From our test we have seen that the cleaning calculation upgrade the log record size well. It lessens the size of document as 72 %. The boundary for the apparatus structured has been seemed in Fig.3

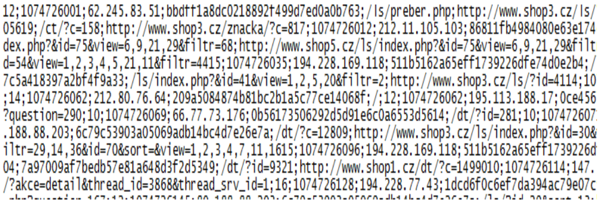


C:\Program Files\MATLAB\R2018a\bin\Wor

Delimiting Character : SEMCOLON

Fig3. Interface tool for Pre-processing

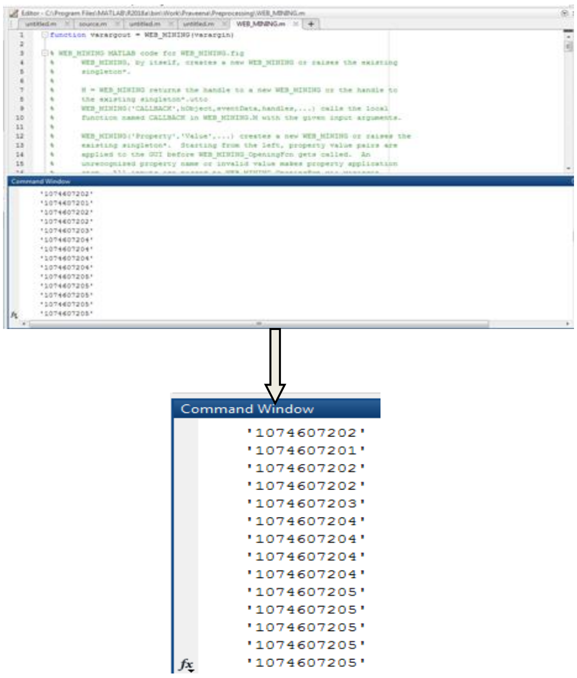
After the data was collected it was directed for processing in the tool created. The yields generated later clicking on each tab illustrated in the proposed effort have been presented. The 'raw web log' data is exposed below,



```
12;107472601;62.245.83.51;bbdff1a8dc0218892f499d7ed0a0b763;/ls/preber.php;http://www.shop3.cz/ls/
05619;/ct/?c=158;http://www.shop3.cz/znacka/?c=817;1074726012;212.11.105.103;86811fb4984080e63e174
dex.php?&id=75&view=6,9,21,29&filtr=68;http://www.shop5.cz/ls/index.php?&id=75&view=6,9,21,29&filt
d=54&view=1,2,3,4,5,21,11&filtr=4415;1074726035;194.228.169.118;511b5162a65eff1739226dfe74d0e2b4;/
7c5a418397a2bf4f9a33;/ls/index.php?&id=41&view=1,2,5,20&filtr=2;http://www.shop3.cz/ls/?id=4114;10
;14;1074726062;212.80.76.64;209a5084874b81bc2b1a5c77ce14068f;/;12;1074726062;195.113.188.17;0ce456
?question=290;10;1074726069;66.77.73.176;0b56173506292d5d91e6c0a6553d5614;/dt/?id=281;10;107472607
.188.88.203;6c79c53903a05069adb14bc4d7e26e7a;/dt/?c=12809;http://www.shop3.cz/ls/index.php?&id=30&
iltr=29,14,36&id=70&sort=&view=1,2,3,4,7,11,1615;1074726096;194.228.169.118;511b5162a65eff1739226d
04;7a97009af7bedb57e81a648d3f2d5349;/dt/?id=9321;http://www.shop1.cz/dt/?c=1499010;107472611 4;147.
/?akce=detail&thread_id=3868&thread_srv_id=1;16;1074726128;194.228.77.43;1dcd6f0c6ef7da394ac79e07c
```

Fig4. Web log data before Pre-processing

Data Cleaning: This tab achieves the cleaning of input data regarding to the algorithm steps conferred in the above Segment.



Editor - C:\Program Files\MATLAB\R2018a\bin\Work\Praveena\Preprocessing\WEB_MINING.m
untitled.m
source.m
untitled.m
untitled.m
WEB_MINING.m
Command Window
Command Window

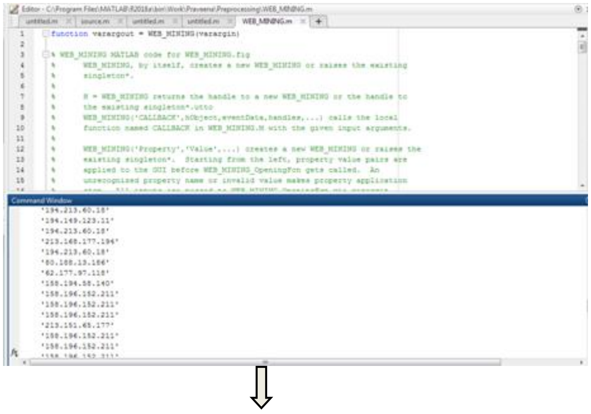
```
function varargout = WEB_MINING(varargin)

% WEB_MINING MATLAB code for WEB_MINING.fig
%      WEB_MINING, by itself, creates a new WEB_MINING or raises the existing
%      singleton*.
%
%      H = WEB_MINING returns the handle to a new WEB_MINING or the handle to
%      the existing singleton*.utto
%      WEB_MINING('CALLBACK',hObject,eventData,handles,...) calls the local
%      function named CALLBACK in WEB_MINING.M with the given input arguments.
%
%      WEB_MINING('Property','Value',...) creates a new WEB_MINING or raises the
%      existing singleton*.  Starting from the left, property value pairs are
%      applied to the GUI before WEB_MINING_OpeningFcn gets called.  An
%      unrecognized property name or invalid value makes property application
```

```
'1074607202'
'1074607201'
'1074607202'
'1074607202'
'1074607203'
'1074607204'
'1074607204'
'1074607204'
'1074607204'
'1074607205'
'1074607205'
'1074607205'
'1074607205'
'1074607205'
```

```
'1074607202'
'1074607201'
'1074607202'
'1074607202'
'1074607203'
'1074607204'
'1074607204'
'1074607204'
'1074607204'
'1074607205'
'1074607205'
'1074607205'
'1074607205'
'1074607205'
```

Fig5. Data Cleaning

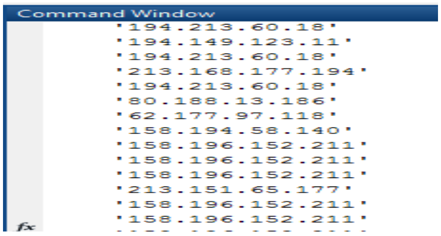
User Identification: When we click on 'user identification' button on our designed tool, we will acquire the users 'IP address'. Operators are the one who stayed the 'web site'. This portion is answerable for recognizing that IP addresses stayed the 'website'. The results exposed have been produced after implementing the algorithm utilized in this research.



Editor - C:\Program Files\MATLAB\R2018a\bin\Work\Praveena\Preprocessing\WEB_MINING.m
untitled.m
source.m
untitled.m
untitled.m
WEB_MINING.m
Command Window

```
function varargout = WEB_MINING(varargin)

% WEB_MINING MATLAB code for WEB_MINING.fig
%      WEB_MINING, by itself, creates a new WEB_MINING or raises the existing
%      singleton*.
%
%      H = WEB_MINING returns the handle to a new WEB_MINING or the handle to
%      the existing singleton*.utto
%      WEB_MINING('CALLBACK',hObject,eventData,handles,...) calls the local
%      function named CALLBACK in WEB_MINING.M with the given input arguments.
%
%      WEB_MINING('Property','Value',...) creates a new WEB_MINING or raises the
%      existing singleton*.  Starting from the left, property value pairs are
%      applied to the GUI before WEB_MINING_OpeningFcn gets called.  An
%      unrecognized property name or invalid value makes property application
```

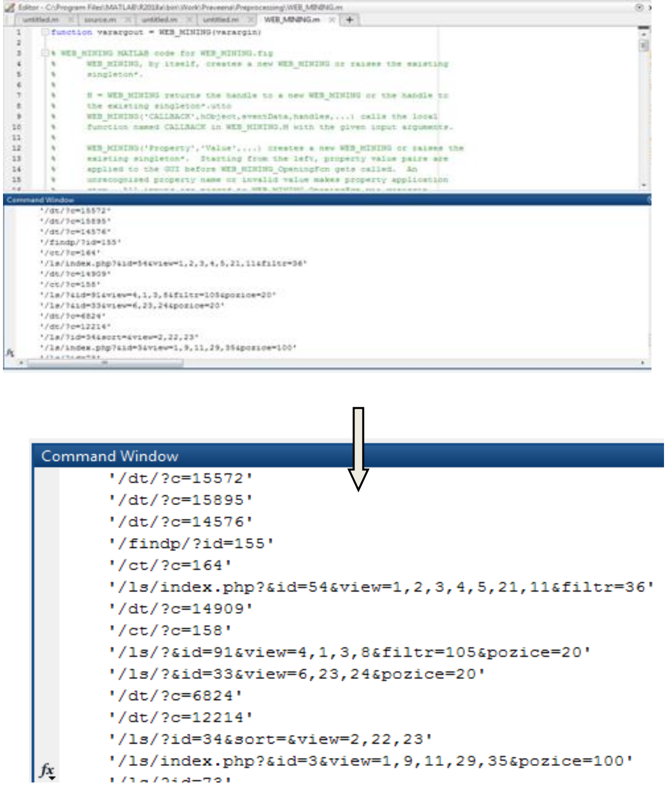
```
'194.213.60.18'
'194.149.123.11'
'194.213.60.18'
'213.168.177.194'
'194.213.60.18'
'80.188.13.186'
'62.177.97.118'
'158.194.58.140'
'158.196.152.211'
'158.196.152.211'
'158.196.152.211'
'213.151.65.177'
'158.196.152.211'
'158.196.152.211'
```



```
'194.213.60.18'
'194.149.123.11'
'194.213.60.18'
'213.168.177.194'
'194.213.60.18'
'80.188.13.186'
'62.177.97.118'
'158.194.58.140'
'158.196.152.211'
'158.196.152.211'
'158.196.152.211'
'213.151.65.177'
'158.196.152.211'
'158.196.152.211'
```

Fig6. User Identification

Session identification: In session identification, the session is started when a client signs onto the site. After the client reorganization, the subsequent stage is 'session identification'. It is the way toward parting the client into the gathering of pages as indicated by the time interim which is for each client discover the session. There are 2 heuristics strategy to discover the meeting. For example, time situated and structure arranged. Right now, have applied structure arranged empirical.



Editor - C:\Program Files\MATLAB\R2018a\bin\Work\Praveena\Preprocessing\WEB_MINING.m
untitled.m
source.m
untitled.m
untitled.m
WEB_MINING.m
Command Window
Command Window

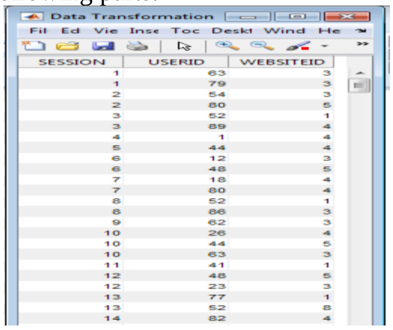
```
function varargout = WEB_MINING(varargin)

% WEB_MINING MATLAB code for WEB_MINING.fig
%      WEB_MINING, by itself, creates a new WEB_MINING or raises the existing
%      singleton*.
%
%      H = WEB_MINING returns the handle to a new WEB_MINING or the handle to
%      the existing singleton*.utto
%      WEB_MINING('CALLBACK',hObject,eventData,handles,...) calls the local
%      function named CALLBACK in WEB_MINING.M with the given input arguments.
%
%      WEB_MINING('Property','Value',...) creates a new WEB_MINING or raises the
%      existing singleton*.  Starting from the left, property value pairs are
%      applied to the GUI before WEB_MINING_OpeningFcn gets called.  An
%      unrecognized property name or invalid value makes property application
```

```
'/dt/?c=15572'
'/dt/?c=15895'
'/dt/?c=14576'
'/findp/?id=155'
'/ct/?c=164'
'/ls/index.php?&id=54&view=1,2,3,4,5,21,11&filtr=36'
'/dt/?c=14909'
'/ct/?c=158'
'/ls/?&id=91&view=4,1,3,8&filtr=105&pozice=20'
'/ls/?&id=33&view=6,23,24&pozice=20'
'/dt/?c=6824'
'/dt/?c=12214'
'/ls/?id=34&sort=&view=2,22,23'
'/ls/index.php?&id=3&view=1,9,11,29,35&pozice=100'
```

```
'/dt/?c=15572'
'/dt/?c=15895'
'/dt/?c=14576'
'/findp/?id=155'
'/ct/?c=164'
'/ls/index.php?&id=54&view=1,2,3,4,5,21,11&filtr=36'
'/dt/?c=14909'
'/ct/?c=158'
'/ls/?&id=91&view=4,1,3,8&filtr=105&pozice=20'
'/ls/?&id=33&view=6,23,24&pozice=20'
'/dt/?c=6824'
'/dt/?c=12214'
'/ls/?id=34&sort=&view=2,22,23'
'/ls/index.php?&id=3&view=1,9,11,29,35&pozice=100'
```

Fig7. Session Identification

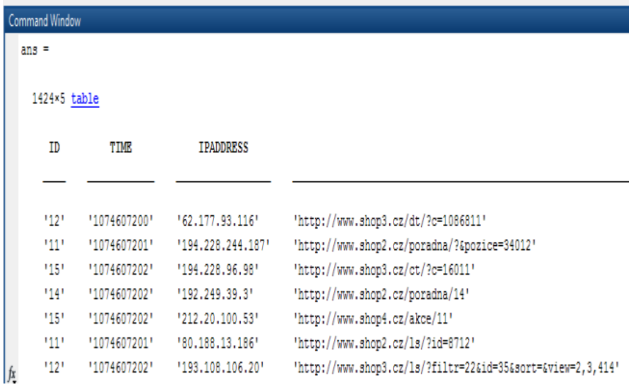
Data transformation: This phase is mostly utilized to transform data to the 'forms' that are apt for 'data mining'. It consists the following parts:



SESSION	USERID	WEBSITEID
1	63	3
1	79	3
2	54	3
2	80	5
3	52	1
3	89	4
4	1	4
5	44	4
6	12	3
6	48	5
7	18	4
7	80	4
8	52	1
8	86	3
9	62	3
10	26	4
10	44	5
10	63	3
11	41	1
12	48	5
12	23	3
13	77	1
13	52	8
14	82	4

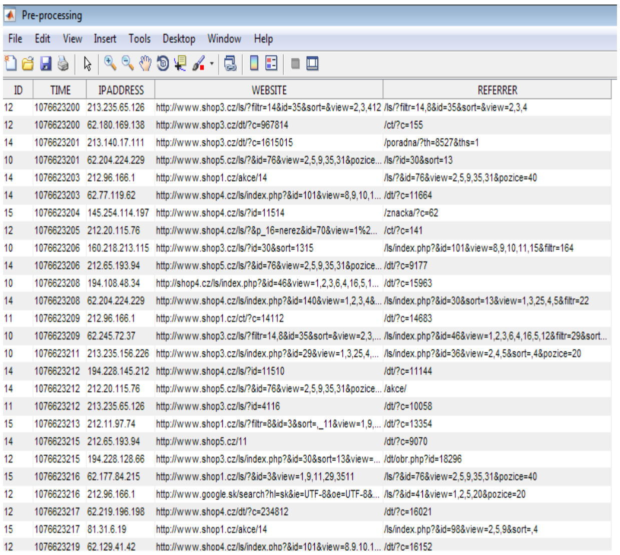
Fig8. Data transformation

Data Formatting: After preprocessing, the output files are accumulated in 'filter log table'. It holds the features about 'user name', 'IP address', 'website address', 'session duration/frequency of visit' that is represented in table 3.1.



```
ans =

  1424×5 table

    ID        TIME           IPADDRESS
   ____    ___________    _______________    ______________________________________________________

   '12'    '1074607200'   '62.177.93.116'     'http://www.shop3.cz/dt/?c=1086811'
   '11'    '1074607201'   '194.228.244.187'   'http://www.shop2.cz/poradna/?&pozice=34012'
   '15'    '1074607202'   '194.228.96.98'     'http://www.shop3.cz/ct/?c=16011'
   '14'    '1074607202'   '192.249.39.3'      'http://www.shop2.cz/poradna/14'
   '15'    '1074607202'   '212.20.100.53'     'http://www.shop4.cz/akce/11'
   '11'    '1074607201'   '80.188.13.186'     'http://www.shop2.cz/ls/?id=8712'
   '12'    '1074607202'   '193.108.106.20'    'http://www.shop3.cz/ls/?filtr=22&id=35&sort=&view=2,3,414'
```

Fig9. Data Formatting



ID	TIME	IPADDRESS	WEBSITE	REFERRER
12	1076623200	213.235.65.126	http://www.shop3.cz/ls/?filtr=14&id=35&sort=&view=2,3,412	/ls/?filtr=14,8&id=35&sort=&view=2,3,4
12	1076623200	62.180.169.138	http://www.shop3.cz/dt/?c=967814	/ct/?c=155
14	1076623201	213.140.17.111	http://www.shop3.cz/dt/?c=1615015	/poradna/?th=8527&ths=1
10	1076623201	62.204.224.229	http://www.shop5.cz/ls/?&id=76&view=2,5,9,35,31&pozice...	/ls/?id=30&sort=13
14	1076623203	212.96.166.1	http://www.shop1.cz/akce/14	/ls/?&id=76&view=2,5,9,35,31&pozice=40
14	1076623203	62.77.119.62	http://www.shop4.cz/ls/index.php?&id=101&view=8,9,10,1...	/dt/?c=11664
15	1076623204	145.254.114.197	http://www.shop4.cz/ls/?id=11514	/znacka/?c=62
12	1076623205	212.20.115.76	http://www.shop4.cz/ls/?&p_16=nerez&id=70&view=1%2...	/ct/?c=141
10	1076623206	160.218.213.115	http://www.shop3.cz/ls/?id=30&sort=1315	/ls/index.php?&id=101&view=8,9,10,11,15&filtr=164
14	1076623206	212.65.193.94	http://www.shop5.cz/ls/?&id=76&view=2,5,9,35,31&pozice...	/dt/?c=9177
10	1076623208	194.108.48.34	http://shop4.cz/ls/index.php?&id=46&view=1,2,3,6,4,16,5,1...	/dt/?c=15963
14	1076623208	62.204.224.229	http://www.shop4.cz/ls/index.php?&id=140&view=1,2,3,4&...	/ls/index.php?&id=30&sort=13&view=1,3,25,4,5&filtr=22
11	1076623209	212.96.166.1	http://www.shop1.cz/ct/?c=14112	/dt/?c=14683
10	1076623209	62.245.72.37	http://www.shop3.cz/ls/?filtr=14,8&id=35&sort=&view=2,3,...	/ls/index.php?&id=46&view=1,2,3,6,4,16,5,12&filtr=29&sort...
10	1076623211	213.235.156.226	http://www.shop3.cz/ls/index.php?&id=29&view=1,3,25,4,...	/ls/index.php?&id=36&view=2,4,5&sort=,4&pozice=20
14	1076623212	194.228.145.212	http://www.shop4.cz/ls/?id=11510	/dt/?c=11144
14	1076623212	212.20.115.76	http://www.shop5.cz/ls/?&id=76&view=2,5,9,35,31&pozice...	/akce/
11	1076623212	213.235.65.126	http://www.shop3.cz/ls/?id=4116	/dt/?c=10058
15	1076623213	212.11.97.74	http://www.shop1.cz/ls/?filtr=8&id=3&sort=,_11&view=1,9,...	/dt/?c=13354
14	1076623215	212.65.193.94	http://www.shop5.cz/11	/dt/?c=9070
12	1076623215	194.228.128.66	http://www.shop3.cz/ls/index.php?&id=30&sort=13&view=...	/dt/obr.php?id=18296
15	1076623216	62.177.84.215	http://www.shop1.cz/ls/?&id=3&view=1,9,11,29,3511	/ls/?&id=76&view=2,5,9,35,31&pozice=40
12	1076623216	212.96.166.1	http://www.google.sk/search?hl=sk&ie=UTF-8&oe=UTF-8&...	/ls/?&id=41&view=1,2,5,20&pozice=20
12	1076623217	62.219.196.198	http://www.shop4.cz/dt/?c=234812	/dt/?c=16021
15	1076623217	81.31.6.19	http://www.shop1.cz/akce/14	/ls/index.php?&id=98&view=2,5,9&sort=,4
12	1076623219	62.129.41.42	http://www.shop4.cz/ls/index.php?&id=101&view=8,9,10,1...	/dt/?c=16152

Fig9. Outcome of preprocessing web log files

Right now, 'web access logs' of our college site were dissected. It has been demonstrated simply that complete number of files decreased considerably after each progression in pre-processing was done. It has been demonstrated simply that cleaning of records outcomes deficient evacuation of superfluous information. 'Clustering' plus 'Neural Network' methods are utilized on the cleaned log record to get high optimistic classified outcomes for web examination. After Hurry up extraction time once users‟ intrigued data is recovered also users‟ got to pages is found from log data.

IV. CONCLUSION

To care web clients in picking up understanding into enormous data on the World Wide Web (WWW), web mining is significant and promising research problematic. The current paper talks round the significance and criticality of 'web log pre-processing'. This research work is proposed to perform data cleaning, client distinguishing proof, meeting recognizable proof and information change. These calculations extricate fields, for example, IP address, user name, website address, session and frequency. This effort offered investigational results of St.Xavier's College of Palayamkottai web server access logs and methods on data preprocessing experimental systems are utilized to clean the construed rough 'web logs'. There are a few procedures and strategies are applied right now to preprocess student access web log files and make them reliable.

REFERENCE

1. V.V.R. Maheswara Rao, V. Valli Kumari, K.V.S.V.N. Raju "An Intelligent System for Web Usage Data Preprocessing" CCSIT 2011, Part I, CCIS 131, pp. 481–490, 2011. © Springer-Verlag Berlin Heidelberg.
2. A V Srinivas "A Survey on Preprocessing of Web-Log Data in Web Usage Mining" International Journal for Modern Trends in Science and Technology Volume: 03, Issue No: 02, February 2017, ISSN: 2455-3778.
3. K.DHARMARAJAN, K.ABIRAMI "Data preprocessing algorithmic approach to identifying user pattern behavior from web server log file" International Journal of Mechanical and Production Engineering Research and Development (IJMPERD), ISSN (P): 2249-6890; ISSN (E): 2249-8001, Vol. 8, Special Issue 3, Dec 2018, 1434-1446, TJPRC Pvt. Ltd.
4. Liu Kewen , "Analysis of Preprocessing Methods for Web Usage Data", in 2012 IEEE International Conference on Measurement, Information and Control, 2012 , p. 383.
5. Neha Goel, Dr. C.K.Jha "Preprocessing Web logs: A Critical phase in Web Usage Mining" IEEE, International Conference on Advances in Computer Engineering and Applications (ICACEA), 2015, 978-1-4673-6911-4/15.
6. Madihah Mohd Saudi, Farida Ridzuan, Member, IAENG, and Hasan Al-Banna Hashim "An Efficient Data Transformation Technique for Web Log" Proceedings of the World Congress on Engineering 2017 Vol IWCE 2017, July 5-7, 2017, London, U.K.
7. P. Dhanalakshmi, K. Ramani, B. Eswara Reddy "The Research of Preprocessing and Pattern Discovery Techniques on Web Log Files" 2016 IEEE 6th International Conference on Advanced Computing (IACC).
8. P. Sukumar, L. Robert, S. Yuvaraj "Review on modern Data Preprocessing techniques in Web usage mining (WUM)" International Conference on Computational Systems and Information Systems for Sustainable Solutions 978-1-5090-1022-6, 2016, IEEE.
9. Michal Munk, Martin Drlík, L'ubomír Benko, Jaroslav Reichel "Quantitative and Qualitative Evaluation of Sequence Patterns Found by Application of Different Educational Data Preprocessing Techniques" 16950497, 23 May 2017, Page(s): 8989 - 9004, IEEE Access (Volume: 5).

AUTHORS PROFILE

A. Dhana Praveena is a research scholar in Mother Teresa Women's University and published a paper on the tools used for wum. She is also working as an Assistant Professor in St. Xavier's College, Playamkottai. She is a member of Board of Studies in St. Mary's College, Thoothukudi.

Dr. V. Selvi is working as an Assistant Professor at Mother Teresa Women's University, Kodaikanal. She is having more than eight years of research experience and guided more than 18 M. Phil scholars and guiding three Ph. D candidates. She is being the member of committee for internal quality assurance cell from 2016 and member of Board of Studies. She is also being the member of College Affiliation Committee. She has published around 25 papers in International Journals.