International Journal of Innovations in Engineering and Technology (IJIET)
http://dx.doi.org/10.21172/ijiet.104.18

An Algorithm for user Identification for Web Usage Mining

Jayanti Mehra[1], R S Thakur[2]
[1,2]*Department of Master of Computer Application, Maulana Azad National Institute of Technology, Bhopal, MP, India*

Abstract- We can see every day web sites are increasing and it is the major problem of web site designing. We can undoubtedly acknowledge how to sites are individually utilized, how we can explore the site, and what number of clients and how much time utilize the site. This is done only through Web Uses Mining. Information hotspots for our web mining reason for existing are customer side treats, web server log, software agent and so on. This paper presents, how web server log information is preprocesses, which incorporates Data cleaning, user identification and Sessionization, path compilation. Once the data is preprocessed it is utilized for finding some useful patterns.
Keywords – Web usage mining, data preprocessing, data cleaning, and user identification.

I. INTRODUCTION

The Internet is an archive of site pages that gives the part of data to the web clients. For web clients the data accessible on web has turned into a fundamental source. On account of these reasons, there is expanding development and many-sided quality of sites accessible on web, the measure of web is vast. A site is the connection the client to organization. The organizations can think about guest's exercises through web examination, and discover the examples. Web mining is comprehensively characterized as finding and investigation of helpful data from the Internet. Web mining partitioned into three sections: Web content Mining, Web structure mining and Web Uses Mining. [5] Web Substance Mining can be as the programmed pursuit and extraction of data and assets accessible from number of locales and on-line databases however web indexes or web arachnids. Web Uses Mining can be as the revelation and examination of access examples of client, through the mining of log documents. The yield of the WUM can be utilized as a part of web personalization, enhancing the framework execution, website adjustment, utilization portrayal and so on.[4] Web log record is a server log document which is a fundamental information sources in Web utilization mining, in which it contain - get to logs of the web server.. The imperative assignment in the WUM is data Preprocessing stage. [9] It comprises of data cleaning, user identification, session identification and path compilation. [3][10]

II. WEB USAGE MINING

Web usage mining is an area of research where Web logs are evaluated and mined to predict user's navigation behavior.[7] Basic techniques of web log mining are association rule mining, clustering, classification etc. are used to discover useful patterns from Web logs. [9][5]

- Collect the log data
- Preprocessing of log data such as data cleaning, User identification, Session Identification, path completion etc
- Log data analysis to discover the useful pattern.
- Discovered patterns evaluation.
- Observance the assessment of discovered patterns.

Different methods of data preprocessing
Data cleaning
User identification
Session identification
Path Completion

III. PREPROCESSING OF WEB USAGE DATA

For the most part in the web uses mining, the preprocessing [4] is considered as an essential advance of web data. As it was recommended in referred paper [1], in the preprocessing incomplete, noisy and fragmented information can expel from data base. Today's, it isn't conceivable to discover great quality information and there is no better result

for mining the greatness information. However, the quality choices rely upon quality information. The copy or missing information may make erroneous result. [11] that is way preprocessing undertaking is mandatory for web log data. [1]

3.1 Data cleaning

Data cleaning [12] is the essential advance for weblog data preprocessing. In the wake of gathering the data, wrong records are expelled from the web log data. Data cleaning is the way toward evacuating the noisy and unessential data which are not expected to the way toward mining. [1] [6]

3.2 User and Session Identification

In the web log, different user sessions can be perceived by user and session identification. [8] Session identification is the strategy for isolating the individual client gets to sign into sessions. [6][14][15]

3.3 Data Collection

We are using raw web log data file from the data repository at http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html. Blow screen shot shows a raw web log file and it is used for our mining process.

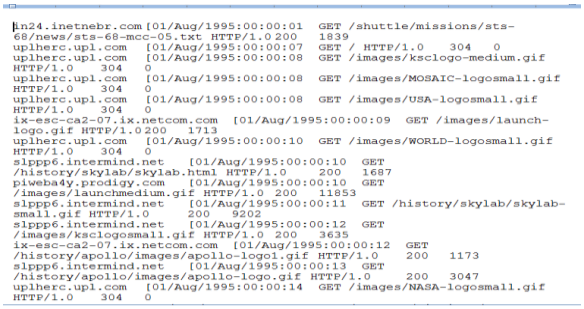
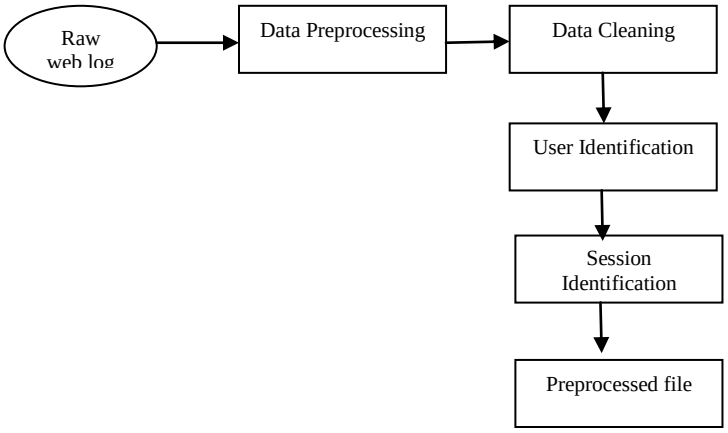


Fig1 Raw web log data

3.4 Implementation work

Our approach that can be given with the help of block diagram –



3.5 Data Cleaning

The initial step of data preprocessing is expelling unnecessary requests from the log records. For the most part, this procedure expels demands, for example, pictures, multimedia documents, page style records, HTTP errors and so forth.

Algorithm 1: Algorithm implemented in java jdk1.7.0. Data Cleaning is proposed to remove inappropriate and missing data from web log files because such data have no use in our analysis. Algorithm also removes failed error status code and eliminates them from further analysis.

Algorithm 1:
Data Cleaning
Input: Raw Log File
Output: cleaned Log file

Algorithm for data cleaning:
Data Cleaning
Info: Log Record Yield:
Cleaned Log Table Algorithm:
1) Open a Log record file
2) Make a table to store log information
3) Open Log Record file
4) Read all fields contain in Log Record
5) For each record in Log Table
On the off chance that (status code='200' and Method='Get')
{
Evacuate fields where '.jpg' or '.jpeg' or '.gif' or '.png' or 'robot.txt' or '.css')
}
Next Record End

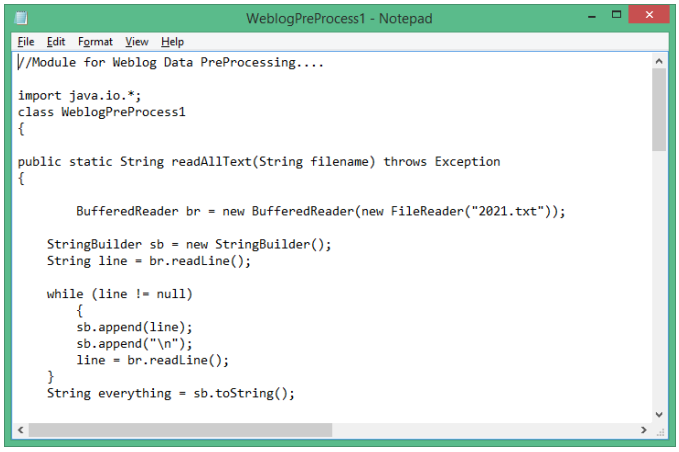
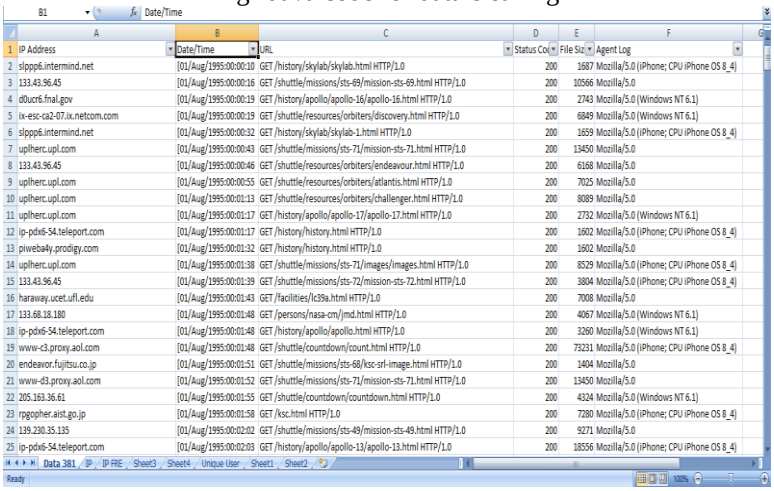


Fig2 Java code for data cleaning



IP Address	Date/Time	URL	Status Co	File Siz	Agent Log
slppp6.intermind.net	[01/Aug/1995:00:00:10	GET /history/skylab/skylab.html HTTP/1.0	200	1687	Mozilla/5.0 (iPhone; CPU iPhone OS 8_4)
133.43.96.45	[01/Aug/1995:00:00:16	GET /shuttle/missions/sts-69/mission-sts-69.html HTTP/1.0	200	10566	Mozilla/5.0
d0ucr6.fnal.gov	[01/Aug/1995:00:00:19	GET /history/apollo/apollo-16/apollo-16.html HTTP/1.0	200	2743	Mozilla/5.0 (Windows NT 6.1)
ix-esc-ca2-07.ix.netcom.com	[01/Aug/1995:00:00:19	GET /shuttle/resources/orbiters/discovery.html HTTP/1.0	200	6849	Mozilla/5.0 (Windows NT 6.1)
slppp6.intermind.net	[01/Aug/1995:00:00:32	GET /history/skylab/skylab-1.html HTTP/1.0	200	1659	Mozilla/5.0 (iPhone; CPU iPhone OS 8_4)
uplherc.upl.com	[01/Aug/1995:00:00:43	GET /shuttle/missions/sts-71/mission-sts-71.html HTTP/1.0	200	13450	Mozilla/5.0
133.43.96.45	[01/Aug/1995:00:00:46	GET /shuttle/resources/orbiters/endeavour.html HTTP/1.0	200	6168	Mozilla/5.0
uplherc.upl.com	[01/Aug/1995:00:00:55	GET /shuttle/resources/orbiters/atlantis.html HTTP/1.0	200	7025	Mozilla/5.0
uplherc.upl.com	[01/Aug/1995:00:01:13	GET /shuttle/resources/orbiters/challenger.html HTTP/1.0	200	8089	Mozilla/5.0
uplherc.upl.com	[01/Aug/1995:00:01:17	GET /history/apollo/apollo-17/apollo-17.html HTTP/1.0	200	2732	Mozilla/5.0 (Windows NT 6.1)
ip-pdx6-54.teleport.com	[01/Aug/1995:00:01:17	GET /history/history.html HTTP/1.0	200	1602	Mozilla/5.0 (iPhone; CPU iPhone OS 8_4)
piweba4y.prodigy.com	[01/Aug/1995:00:01:32	GET /history/history.html HTTP/1.0	200	1602	Mozilla/5.0
uplherc.upl.com	[01/Aug/1995:00:01:38	GET /shuttle/missions/sts-71/images/images.html HTTP/1.0	200	8529	Mozilla/5.0 (iPhone; CPU iPhone OS 8_4)
133.43.96.45	[01/Aug/1995:00:01:39	GET /shuttle/missions/sts-72/mission-sts-72.html HTTP/1.0	200	3804	Mozilla/5.0 (iPhone; CPU iPhone OS 8_4)
haraway.ucet.ufl.edu	[01/Aug/1995:00:01:43	GET /facilities/lc39a.html HTTP/1.0	200	7008	Mozilla/5.0
133.68.18.180	[01/Aug/1995:00:01:48	GET /persons/nasa-cm/jmd.html HTTP/1.0	200	4067	Mozilla/5.0 (Windows NT 6.1)
ip-pdx6-54.teleport.com	[01/Aug/1995:00:01:48	GET /history/apollo/apollo.html HTTP/1.0	200	3260	Mozilla/5.0 (Windows NT 6.1)
www-c3.proxy.aol.com	[01/Aug/1995:00:01:48	GET /shuttle/countdown/count.html HTTP/1.0	200	73231	Mozilla/5.0 (iPhone; CPU iPhone OS 8_4)
endeavor.fujitsu.co.jp	[01/Aug/1995:00:01:51	GET /shuttle/missions/sts-68/ksc-srl-image.html HTTP/1.0	200	1404	Mozilla/5.0
www-d3.proxy.aol.com	[01/Aug/1995:00:01:52	GET /shuttle/missions/sts-71/mission-sts-71.html HTTP/1.0	200	13450	Mozilla/5.0
205.163.36.61	[01/Aug/1995:00:01:55	GET /shuttle/countdown/countdown.html HTTP/1.0	200	4324	Mozilla/5.0 (Windows NT 6.1)
rgopher.aist.go.jp	[01/Aug/1995:00:01:58	GET /ksc.html HTTP/1.0	200	7280	Mozilla/5.0 (iPhone; CPU iPhone OS 8_4)
139.230.35.135	[01/Aug/1995:00:02:02	GET /shuttle/missions/sts-49/mission-sts-49.html HTTP/1.0	200	9271	Mozilla/5.0
ip-pdx6-54.teleport.com	[01/Aug/1995:00:02:03	GET /history/apollo/apollo-13/apollo-13.html HTTP/1.0	200	18556	Mozilla/5.0 (iPhone; CPU iPhone OS 8_4)

Fig3 Cleaned web log data

http://dx.doi.org/10.21172/ijiet.104.18

IV. USER IDENTIFICATION

After data cleaning we can identifies the user from the database table. [2] In our approach a user is identified as follows:

We can use some information about IP address and client: User's browser and operating system versions are logged into User agent field. The following algorithm identifies a unique user. [6][8]

Input: Cleaned web log file
Output: Total unique user
Start
For every entry in log file
If (distinct IP)
New User
Else
Distinct user agent
New user
End

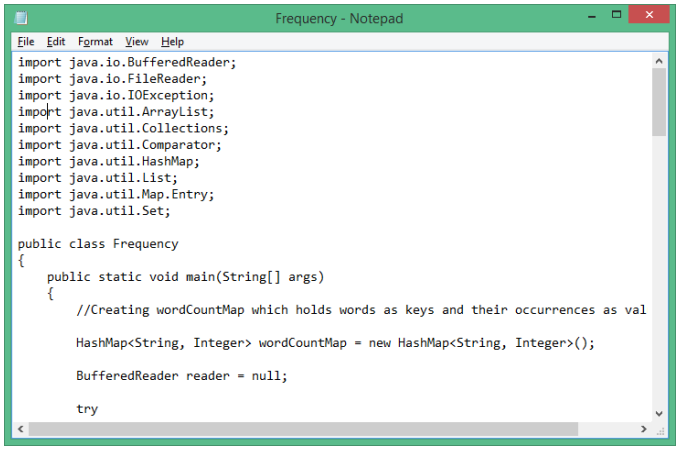
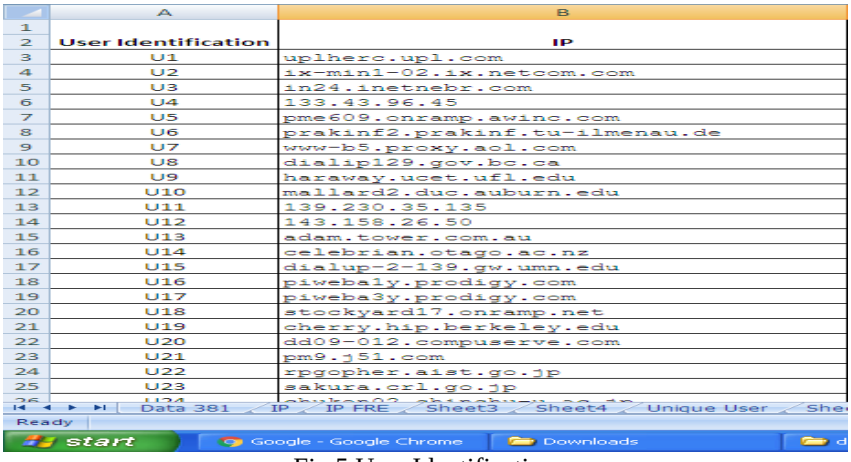


Fig 4 Java code for user Identification



	A	B
1		
2	User Identification	IP
3	U1	uplherc.upl.com
4	U2	ix-min1-02.ix.netcom.com
5	U3	in24.inetnebr.com
6	U4	133.43.96.45
7	U5	pme609.onramp.awinc.com
8	U6	prakinf2.prakinf.tu-ilmenau.de
9	U7	www-b5.proxy.aol.com
10	U8	dialip129.gov.bc.ca
11	U9	haraway.ucet.ufl.edu
12	U10	mallard2.duc.auburn.edu
13	U11	139.230.35.135
14	U12	143.158.26.50
15	U13	adam.tower.com.au
16	U14	celebrian.otago.ac.nz
17	U15	dialup-2-139.gw.umn.edu
18	U16	piweba1y.prodigy.com
19	U17	piweba3y.prodigy.com
20	U18	stockyard17.onramp.net
21	U19	cherry.hip.berkeley.edu
22	U20	dd09-012.compuserve.com
23	U21	pm9.j51.com
24	U22	rpgopher.aist.go.jp
25	U23	sakura.crl.go.jp

Fig 5 User Identification

For the analysis we use matlab software for performing k-means clustering to identifies the group and we found:
[cidx3,cmeans3,sumd3] = kmeans(meas,2,'replicates',5,'display','final');

Group of same user and distinct user matlab window

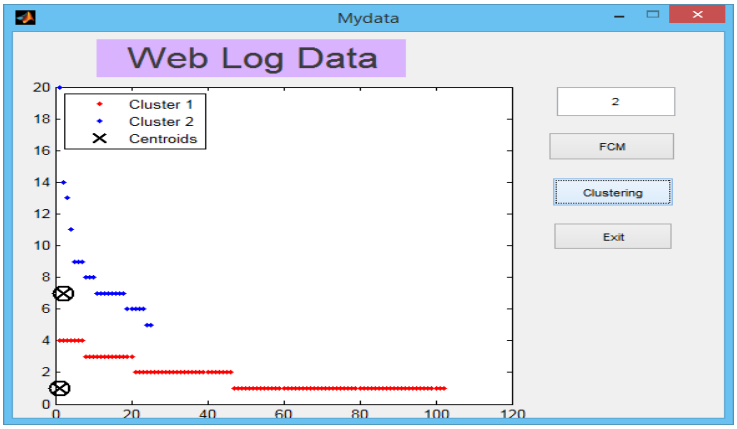


Table 1: Processed log Dataset

Label	Processed Value
Total records	1727
After data cleaning	380
Total User	128
Same user	71
Distinct User	56

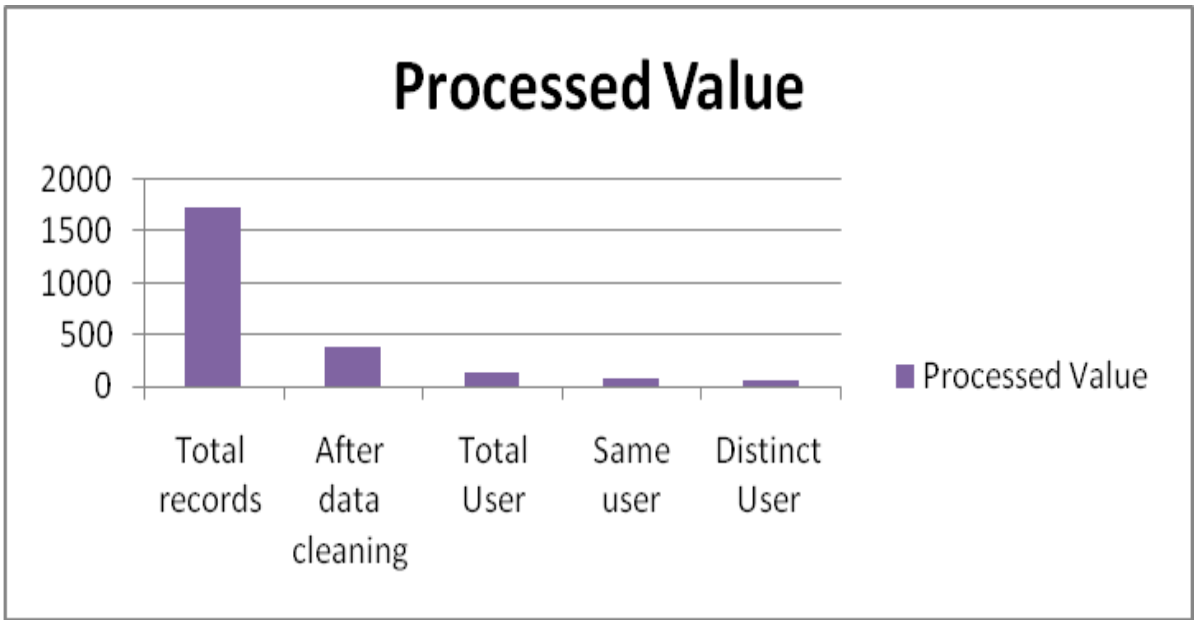


Fig 7 Unique Users Identified as of the cleaned web log file

V. CONCLUSION

In this paper we have talked about data cleaning and distinct user identification procedure to improve the real advance of web log mining, Through user identification we can likewise discover distinct user in light of their went to session time. These procedures create more exact outcome. After total the preprocessing assignment we can without much of a stretch customized sites, build up the outline of Website pages. Future work is this we can done whole procedure of WUM. An entire technique covering, for example, pattern discovery and pattern analysis can be executed on preprocessed data.

VI. REFERENCES

[1] J.Umarani,and K.Karpagam, "Investigation of User Identification Methods in Pre-Processing Phase of Web Usage Mining", International Journal of Engineering Science and Computing, August 2016 Volume 6 issue 8.

[2] Priyanka Patel et al,"A Review on User Session Identification through Web Server Log"(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (1), 2014, 146-148.

[3] Michal Munka, Jozef Kapustaa, Peter Švecaolume ," Data Preprocessing Evaluation for Web Log Mining: Reconstruction of Activities of a Web Visitor " International Conference on Computational Science, ICCS 2010

[4] Chaitra L Mugali et al," Pre-Processing and Analysis of Web Server Logs",

[5] International Journal of Innovative Research in Advanced Engineering (IJIRAE), ISSN: 2349-2163 Issue 8, Volume 2 (August 2015).

[6] Sheetal A. et al," Efficient Preprocessing technique using Web log mining", International Journal of Advancements in Research & Technology, Volume 1, Issue6, November-2012 1 ISSN 2278-7763.

[7] Manisha Valera et al,"A Step up in Data Cleaning and User identification of Preprocessing on Web Usage data", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3, Issue 12, December 2014.

[8] Harmit kaur et al," A Survey of Preprocessing Method for Web Usage Mining User Behavior Understanding", international journal of computers and communications, Volume 8, 2014.

[9] Surbhi Anand," An Efficient Algorithm for Data Cleaning of Log File using File Extensions", International Journal of Computer Applications (0975 –888), Volume 48– No.8, June 2012.

[10] Yadav et al," Algorithms for Web Log Data: WUM Pre-Processing phase", Mansi International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181 Vol. 3 Issue12, December-2014.

[11] Priyanka Verma et al ," Web Usage mining framework for Data Cleaning and IP address Identification", August 2014.

[12] Sujith Jayaprakash et al," A Comprehensive Survey on Data Preprocessing Methods in Web Usage Minning", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3), 2015, 3170-3174.

[13] S. Umamaheswari et al," Algorithm for Tracing Visitors' On-Line Behaviors for Effective Web Usage Mining", International Journal of Computer Applications (0975 – 8887) Volume 87 – No.3, February 2014.

[14] Ketan D. Patel et al, "Preprocessing on Web Server Log Data for Web Usage Pattern Discovery", International Journal of Computer Applications (0975 – 8887) Volume 165 – No.10, May 2017.

[15] V.Chitraa et al," A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing, International Journal of Computer Applications (0975 – 8887) Volume 34– No.9, November 2011.

[16] Zakaria Suliman Zubi et al," Using Web Logs Dataset via Web Mining for user Behavior Understanding", International journal of computers and communications, Volume 8, 2014