Efficient Log Mining from Web Server Using Clustering Technique

Ripal Patel[1], Mr. Krunal Panchal[2], Mr. Dushyantsinh Rathod[3]

[1]M.E., [2,3]Assistant Professor, [1,2,3]computer Engineering Department,
[1,2]L J Institute of Engineering and Technology Ahmedabad, Gujarat, India

Abstract— In this, Web Usage Mining is used to discover interesting patterns in accesses to various Web pages within the Web space associated with a particular server. The Web Usage Mining architecture is divided into two main parts- the first part includes preprocessing and data integration components. The second part includes the largely domain independent application of finally Cleaned data and database file generation. This research work will take less time compare to the existing algorithms. This research work is mainly composed of dynamic web log preprocessing for mined in different applications.

Keywords: Log Files , Web Log Pre-processing, Data Mining, Web Mining, Web Content Mining, Web Usage Mining, Web Structure Mining

I. INTRODUCTION

In dynamic systems such as the Internet, it is a common practice to periodically record samples of activity. Those samples are then used to characterize the activity in the system and to evaluate new mechanisms to be used in this system. It is called Log files. Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications. Usage data captures the identity or origin of web users along with their browsing behaviour at a web site.

II. OBJECTIVE

Input raw web accessing log file. Take user choice for normal, multimedia, graphics or e-commerce applications. Merge Mining data from web log file. Read raw web log file and remove logs according to user selection to make intermediate file. Generate more structured and easily readable database file. Use clustering technique of data mining on pre-processed data and do the analysis.

III. TYPES OF WEB MINING

Web mining is a technique to analyze the online Web contents, navigate between various Web sites and perform transaction of data across the Web. Web mining can be broadly divided into three distinct categories, according to the kinds of data to be mined. Figure 2 shows the taxonomy.

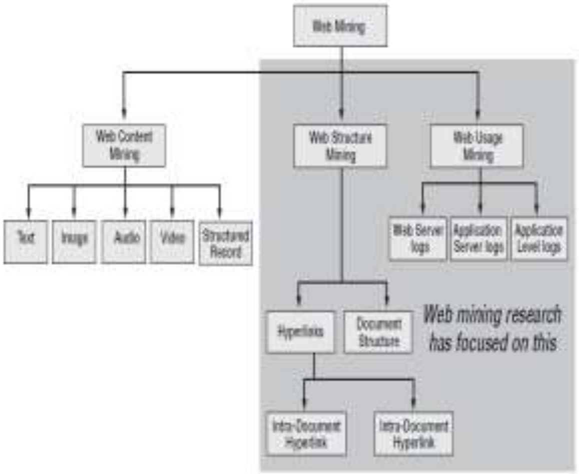


Fig 2: Types of Web Mining[7]

Web Content Mining

Web content mining discovers information or knowledge from millions of sources across the Web. In web content mining, patterns are extracted from online sources such as HTML files, text documents, images, e-books or e-mail messages. The concept of Web content mining is far wider than searching for any specific term or only keyword extraction or some simple statistics of words and phrases in documents. For example, a tool that performs web content mining can summarize a web page so that you need not read the complete document and save your time and energy.

Web Structure Mining The structure of a typical web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Web structure mining is the process of discovering structure information from the web. This can be further divided into two kinds based on the kind of structure information used.

Web Usage Mining [7]

Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications (Srivastava, Cooley, Deshpande, and Tan 2000). Usage data captures the identity or origin of web users along with their browsing behaviour at a web site. Web usage mining itself can be classified further depending on the kind of usage data considered:

A. Web Server Data

User logs are collected by the web server and typically include IP address, page reference and access time.

B. Application Server Data

Commercial application servers such as Weblogic,1,2 StoryServer,3 have significant features to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.

C. Application Level Data

New kinds of events can be defined in an application, and logging can be turned on for them — generating histories of these events. It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the above the categories.

IV. PROBLEM STATEMENT

Issues to be considered by this study; Single log file is used, not customizable, requires more time to perform single process on different log files, less accurate
To overcome the above issues the following solutions are proposed; Use different log file formats in single process. To customize data checkboxes are used for user selection. To reduce time single process can done on different log file format in one time. To increase accuracy k-means clustering is used which gives OUTLIERS.

V. LITERATURE REVIEW

Table 3: Comparison of Literature Survey

Sr No	Description	Approach	Pros	Cons
1	Extracting Knowledge from Web Server Logs Using Web Usage Mining	Web Usage Mining process, W3C log file format	Uses unstructured data	Worked on single log file format W3C
2	An Efficient Web Mining Algorithm for Web Log Analysis: E-Web Miner	Improved AprioriAll algorithm	Better log analysis, number of database scanning get reduced	Uses single log file format
3	Design and Implementation of WEB Log Mining	XML and relational database	Simple and practical approach ,helps to realize the further mining for Web log data from specific time quantum	Worked on single log file format, also increases transaction

			and specific user	
4	Optimized Data Pre-processing Technology for Web Log Mining	User identification based on referred page, session restructuring algorithm	Improvement of the technology betters the quality of data pre-processing results	Used only IIS log file format and more database transaction
5	Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm	Association rule mining algorithm	Improves performance of mining system	Uses only one log file format for mining process

VI. PROPOSED SOLUTION

Steps:
Algorithm 1, we can read web server log data line by line and store into a data source.
Algorithm 2, we can mining and integrate multiple data source.
Algorithm 3, we can remove irrelevant data from data source.
Algorithm 4, we can generate clusters.
In this system it is working for all log file format and final records will be stored into the data source like Data table, Data Set etc.

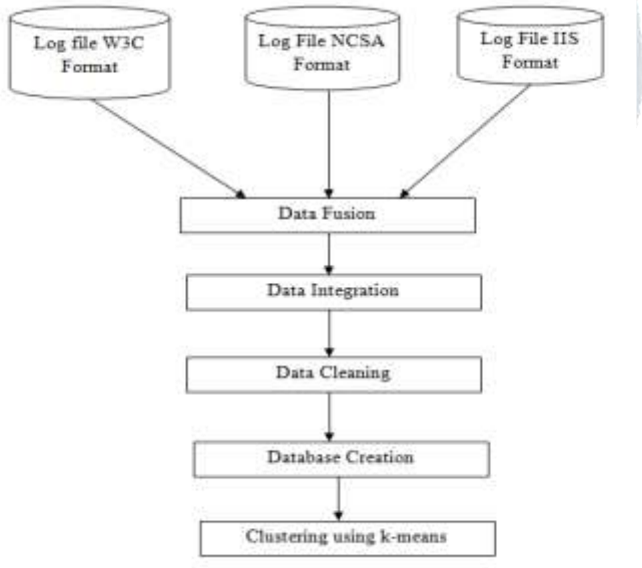


Fig. 3: Framework of proposed model

Clustering steps:

1. Fill gridview with all data from database
2. select multiple client IP address from same server IP address by using for loop to count each row from gridview
3. find selected rows by using checkbox whether the rows is selected or not
4. if rows found selected then take rows id and check using IF condition that client IP address is from the same server IP address
5. Then create new datarow in datatable for creating a cluster 1
6. add all selected rows in datatable
7. if found different server IP then it becomes OUTLIER
8. bind datatable to gridview.

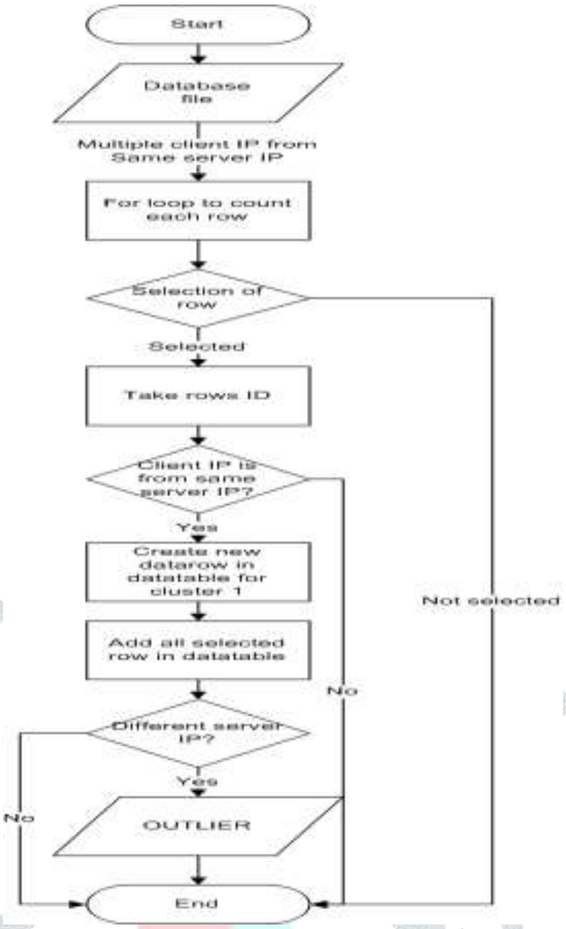

Start
Database file
Multiple client IP from Same server IP
For loop to count each row
Selection of row
Selected
Take rows ID
Client IP is from same server IP?
Yes
Create new datarow in datatable for cluster 1
Add all selected row in datatable
Different server IP?
Yes
OUTLIER
No
No
Not selected
End

Fig. 4: Framework of Clustering Technique

VII. IMPLEMENTATION WORK

Interface for data selection

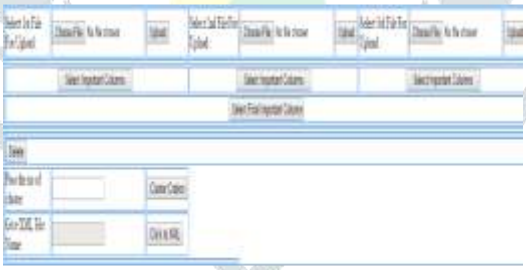
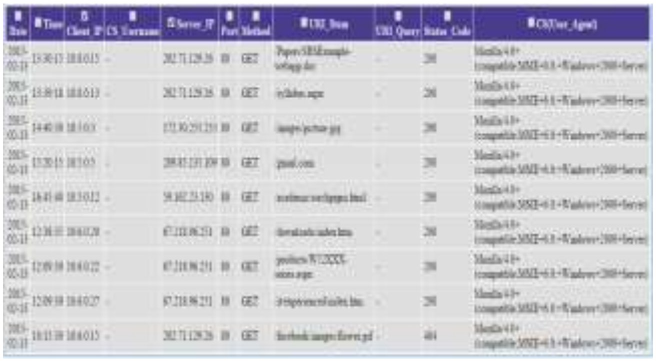

Select 1st File For Upload
Choose File No file chosen
Upload
Select 2nd File For Upload
Choose File No file chosen
Upload
Select 3rd File For Upload
Choose File No file chosen
Upload
Select Important Columns
Select Important Columns
Select Important Columns
Select Final Important Columns
Delete
Provide no of cluster
Cluster Create
Get in XML File Format
Click to XML

Fig 5: Interface for Data Selection

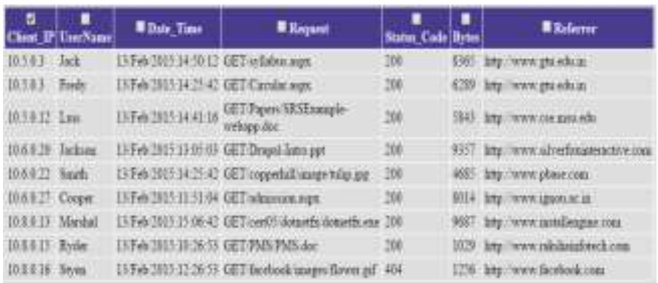
In this you can select 3 different types of log files and perform integration of them and it gives pre-processed file.

Selection of Columns from Log Files



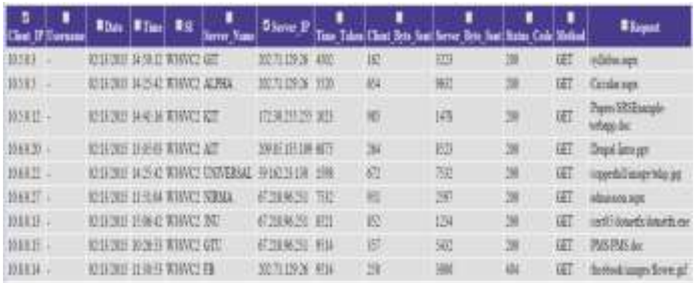
Date	Time	Client_IP	CS_Username	Server_IP	Port	Method	URI_Stem	URI_Query	Status_Code	CS(User_Agent)
2015-05-18	13:30:15	10.0.0.15	-	202.71.129.26	80	GET	Papers/SBSExample-webapp.doc	-	200	Mozilla/4.0+(compatible;MSIE+6.0;+Windows+2000+Server)
2015-05-18	13:30:18	10.0.0.13	-	202.71.129.26	80	GET	syllabus.aspx	-	200	Mozilla/4.0+(compatible;MSIE+6.0;+Windows+2000+Server)
2015-05-18	14:40:18	10.1.0.3	-	172.30.251.231	80	GET	images/picture.jpg	-	200	Mozilla/4.0+(compatible;MSIE+6.0;+Windows+2000+Server)
2015-05-18	13:30:15	10.1.0.5	-	209.85.231.109	80	GET	gmail.com	-	200	Mozilla/4.0+(compatible;MSIE+6.0;+Windows+2000+Server)
2015-05-18	16:41:48	10.1.0.12	-	58.102.23.180	80	GET	[illegible]	-	200	Mozilla/4.0+(compatible;MSIE+6.0;+Windows+2000+Server)
2015-05-18	12:38:33	10.0.0.28	-	67.218.96.231	80	GET	downloads/index.htm	-	200	Mozilla/4.0+(compatible;MSIE+6.0;+Windows+2000+Server)
2015-05-18	12:09:18	10.0.0.22	-	67.218.96.231	80	GET	products/W12XXX-series.aspx	-	200	Mozilla/4.0+(compatible;MSIE+6.0;+Windows+2000+Server)
2015-05-18	12:09:18	10.0.0.27	-	67.218.96.231	80	GET	[illegible]	-	200	Mozilla/4.0+(compatible;MSIE+6.0;+Windows+2000+Server)
2015-05-18	18:11:18	10.0.0.15	-	202.71.129.26	80	GET	facebook/images/flower.gif	-	404	Mozilla/4.0+(compatible;MSIE+6.0;+Windows+2000+Server)

Fig 6: Column Selection from Extended W3C Log File



☑ Client_IP	■ UserName	■ Date_Time	■ Request	■ Status_Code	■ Bytes	■ Referrer
10.5.0.3	Jack	13/Feb/2015:14:50:12	GET/syllabus.aspx	200	5365	http://www.gtu.edu.in
10.5.0.3	Fredy	13/Feb/2015:14:25:42	GET/Circular.aspx	200	6289	http://www.gtu.edu.in
10.5.0.12	Lou	13/Feb/2015:14:41:16	GET/Papers/SRSExample-webapp.doc	200	5843	http://www.cse.msu.edu
10.6.0.20	Jackson	13/Feb/2015:13:05:03	GET/Drupal-Intro.ppt	200	9357	http://www.silverfoxinteractive.com
10.6.0.22	Sarah	13/Feb/2015:14:25:42	GET/copperhill/images/tulip.jpg	200	4685	http://www.pbase.com
10.6.0.27	Cooper	13/Feb/2015:11:51:04	GET/admission.aspx	200	8014	http://www.ignou.ac.in
10.8.0.13	Marshal	13/Feb/2015:15:06:42	GET/cert05/dotnetfx/dotnetfx.exe	200	9687	http://www.installengine.com
10.8.0.15	Ryder	13/Feb/2015:10:26:53	GET/PMS/PMS.doc	200	1029	http://www.rahulsinfotech.com
10.8.0.16	Steyen	13/Feb/2015:12:26:53	GET/facebook/images/flower.gif	404	1256	http://www.facebook.com

Fig 7: Column Selection from NCSA Log File

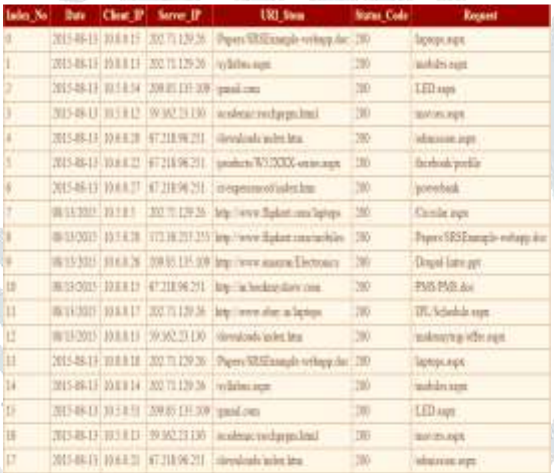


☐ Client_IP	■ Username	■ Date	■ Time	■ SI	■ Server_Name	☐ Server_IP	■ Time_Taken	■ Client_Byts_Sent	■ Server_Byts_Sent	■ Status_Code	■ Method	■ Request
10.5.0.3	-	02/13/2015	14:50:12	W3SVC2	GIT	202.71.129.26	4302	182	3223	200	GET	syllabus.aspx
10.5.0.3	-	02/13/2015	14:25:42	W3SVC2	ALPHA	202.71.129.26	3120	454	9632	200	GET	Circular.aspx
10.5.0.12	-	02/13/2015	14:41:16	W3SVC2	KIT	172.30.255.255	3023	985	1478	200	GET	Papers/SRSExample-webapp.doc
10.6.0.20	-	02/13/2015	13:05:03	W3SVC2	AIT	209.85.135.109	4673	284	1523	200	GET	Drupal-Intro.ppt
10.6.0.22	-	02/13/2015	14:25:42	W3SVC2	UNIVERSAL	59.162.23.130	2598	672	7532	200	GET	copperhill/images/tulip.jpg
10.6.0.27	-	02/13/2015	11:51:04	W3SVC2	NIRMA	67.218.96.251	7512	951	2587	200	GET	admission.aspx
10.8.0.13	-	02/13/2015	15:06:42	W3SVC2	JNU	67.218.96.251	8121	852	1234	200	GET	cert05/dotnetfx/dotnetfx.exe
10.8.0.15	-	02/13/2015	10:26:53	W3SVC2	GTU	67.218.96.251	9114	157	5402	200	GET	PMS/PMS.doc
10.8.0.14	-	02/13/2015	11:30:53	W3SVC2	FB	202.71.129.26	8514	258	3886	404	GET	facebook/images/flower.gif

Fig 8: Column Selection from IIS Log File

Columns are selected by giving check to the box. You can select columns in three log files and get necessary data to process.

Clean Data

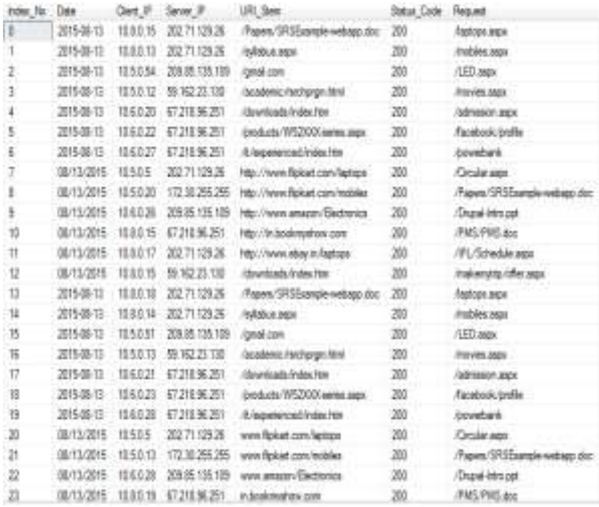


Index_No	Date	Client_IP	Server_IP	URI_Stem	Status_Code	Request
0	2015-08-13	10.8.0.15	202.71.129.26	Papers/SRSExample-webapp.doc	200	laptops.aspx
1	2015-08-13	10.8.0.13	202.71.129.26	syllabus.aspx	200	mobiles.aspx
2	2015-08-13	10.5.0.54	209.85.135.109	gmail.com	200	LED.aspx
3	2015-08-13	10.5.0.12	59.162.23.130	academic/techprgm.html	200	movies.aspx
4	2015-08-13	10.6.0.20	67.218.96.251	downloads/index.htm	200	admission.aspx
5	2015-08-13	10.6.0.22	67.218.96.251	products/WS2XXX-series.aspx	200	facebook/profile
6	2015-08-13	10.6.0.27	67.218.96.251	it/experienced/index.htm	200	powerbank
7	08/13/2015	10.5.0.5	202.71.129.26	http://www.flipkart.com/laptops	200	Circular.aspx
8	08/13/2015	10.5.0.20	172.30.255.255	http://www.flipkart.com/mobiles	200	Papers/SRSExample-webapp.doc
9	08/13/2015	10.6.0.26	209.85.135.109	http://www.amazon/Electronics	200	Drupal-Intro.ppt
10	08/13/2015	10.8.0.15	67.218.96.251	http://in.bookmyshow.com	200	PMS/PMS.doc
11	08/13/2015	10.8.0.17	202.71.129.26	http://www.ebay.in/laptops	200	IPL/Schedule.aspx
12	08/13/2015	10.8.0.15	59.162.23.130	downloads/index.htm	200	makemytrip/offer.aspx
13	2015-08-13	10.8.0.18	202.71.129.26	Papers/SRSExample-webapp.doc	200	laptops.aspx
14	2015-08-13	10.8.0.14	202.71.129.26	syllabus.aspx	200	mobiles.aspx
15	2015-08-13	10.5.0.51	209.85.135.109	gmail.com	200	LED.aspx
16	2015-08-13	10.5.0.13	59.162.23.130	academic/techprgm.html	200	movies.aspx
17	2015-08-13	10.6.0.21	67.218.96.251	downloads/index.htm	200	admission.aspx

Fig 9: Clean Data

After selecting the unnecessary data you can get the clean data to further mining process. Figure shows the clean data after pre-processing. You can get the pre-processed data.

Database creation

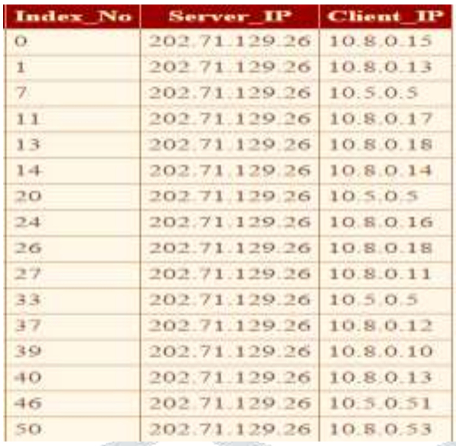


Index_No	Date	Client_IP	Server_IP	URI_Stem	Status_Code	Request
0	2015-08-13	10.0.0.15	202.71.129.26	/Papers/SRSExample-webapp.doc	200	/laptops.aspx
1	2015-08-13	10.0.0.13	202.71.129.26	/syllabus.aspx	200	/mobiles.aspx
2	2015-08-13	10.5.0.54	209.85.135.109	/gmail.com	200	/LED.aspx
3	2015-08-13	10.5.0.12	59.162.23.130	/academic/techprgm.html	200	/movies.aspx
4	2015-08-13	10.6.0.20	67.218.96.251	/downloads/index.htm	200	/admission.aspx
5	2015-08-13	10.6.0.22	67.218.96.251	/products/WS2XXX-series.aspx	200	/facebook/profile
6	2015-08-13	10.6.0.27	67.218.96.251	/it/experienced/index.htm	200	/powerbank
7	08/13/2015	10.5.0.5	202.71.129.26	Http://www.flipkart.com/laptops	200	/Circular.aspx
8	08/13/2015	10.5.0.20	172.30.255.255	Http://www.flipkart.com/mobiles	200	/Papers/SRSExample-webapp.doc
9	08/13/2015	10.6.0.26	209.85.135.109	Http://www.amazon/Electronics	200	/Drupal-Intro.ppt
10	08/13/2015	10.0.0.15	67.218.96.251	Http://in.bookmyshow.com	200	/PMS/PMS.doc
11	08/13/2015	10.0.0.17	202.71.129.26	Http://www.ebay.in/laptops	200	/IPL/Schedule.aspx
12	08/13/2015	10.0.0.15	59.162.23.130	/downloads/index.htm	200	/makemytrip/offer.aspx
13	2015-08-13	10.0.0.18	202.71.129.26	/Papers/SRSExample-webapp.doc	200	/laptops.aspx
14	2015-08-13	10.0.0.14	202.71.129.26	/syllabus.aspx	200	/mobiles.aspx
15	2015-08-13	10.5.0.51	209.85.135.109	/gmail.com	200	/LED.aspx
16	2015-08-13	10.5.0.13	59.162.23.130	/academic/techprgm.html	200	/movies.aspx
17	2015-08-13	10.6.0.21	67.218.96.251	/downloads/index.htm	200	/admission.aspx
18	2015-08-13	10.6.0.23	67.218.96.251	/products/WS2XXX-series.aspx	200	/facebook/profile
19	2015-08-13	10.6.0.28	67.218.96.251	/it/experienced/index.htm	200	/powerbank
20	08/13/2015	10.5.0.5	202.71.129.26	www.flipkart.com/laptops	200	/Circular.aspx
21	08/13/2015	10.5.0.13	172.30.255.255	www.flipkart.com/mobiles	200	/Papers/SRSExample-webapp.doc
22	08/13/2015	10.6.0.28	209.85.135.109	www.amazon/Electronics	200	/Drupal-Intro.ppt
23	08/13/2015	10.0.0.19	67.218.96.251	in.bookmyshow.com	200	/PMS/PMS.doc

Fig 10: Database file

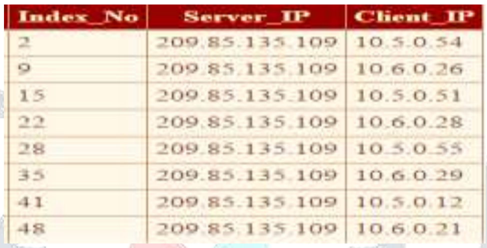
You can create this file using xls file which is created from the clean data or we can say pre-processed data. This file is used in the mining algorithm which is useful to get knowledge for further analysis.

4.1.1 Clusters



Index_No	Server_IP	Client_IP
0	202.71.129.26	10.8.0.15
1	202.71.129.26	10.8.0.13
7	202.71.129.26	10.5.0.5
11	202.71.129.26	10.8.0.17
13	202.71.129.26	10.8.0.18
14	202.71.129.26	10.8.0.14
20	202.71.129.26	10.5.0.5
24	202.71.129.26	10.8.0.16
26	202.71.129.26	10.8.0.18
27	202.71.129.26	10.8.0.11
33	202.71.129.26	10.5.0.5
37	202.71.129.26	10.8.0.12
39	202.71.129.26	10.8.0.10
40	202.71.129.26	10.8.0.13
46	202.71.129.26	10.5.0.51
50	202.71.129.26	10.8.0.53

Fig 11: Cluster 1



Index_No	Server_IP	Client_IP
2	209.85.135.109	10.5.0.54
9	209.85.135.109	10.6.0.26
15	209.85.135.109	10.5.0.51
22	209.85.135.109	10.6.0.28
28	209.85.135.109	10.5.0.55
35	209.85.135.109	10.6.0.29
41	209.85.135.109	10.5.0.12
48	209.85.135.109	10.6.0.21

Fig 12: Cluster 2

Above all figure shown the clusters created from the database file. These clusters are defined mined data from the database.

VIII. COMPARATIVE RESULT

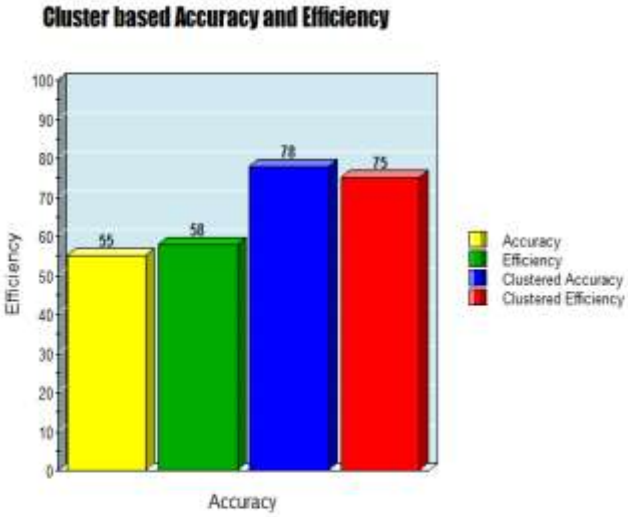


Fig 13: Comparative result

This graph shows the result of comparison between data with cluster and without cluster. It clearly shows that accuracy and efficiency both are comparatively high. From this graph or chart we conclude that with the use of clustering technique we get the better accuracy and efficiency which are the parameters to focus in this research work.

IX. CONCLUTION AND FUTURE WORK

This Research work includes pre-processing phase and web usage mining which can be utilized in industry and application oriented system. We uses customized web log pre-processing rather than traditional approach which may reduces size of raw web log file. Improvement will show in execution time and accuracy. This research work gives customized data from multiple data source so it increases performance of web server and fast analysis of data. It used k-means clustering which gives mined data to analyse. The results of mining can be used to improve the website design and increase satisfaction which helps in various applications.

In future, association rule mining can used to mine the data from the pre-processed data.

References

[1] Eltahir. M.A. , Dafa-Alla, "Extracting knowledge from web server logs using web usage mining", IEEE Aug, 2013, pp.413-417, ISBN: 978-1-4673-6231-3

[2] Yadav, M.P. , Keserwani, P.K. , "An efficient web mining algorithm for Web Log analysis: E-Web Miner", IEEE March, 2012, pp.607-613, ISBN: 978-1-4577-0694-3

[3] Xianjun Ni , "Design and Implementation of WEB Log Mining System", IEEE Jan, 2009, pp.425-427, ISBN: 978-1-4244-3334-6

[4] Ling Zheng, Hui Gui , "Optimized data preprocessing technology for web log mining", IEEE June, 2010, pp.V1-19 – V1-22, ISBN: 978-1-4244-7164-5

[5] Mohd Helmy Abd Wahab, Mohd Norzali Haji Mohd, Hafizul Fahri Hanafi, Mohamad Farhan, Mohamad Mohsin, "Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm", World Academy of Science, Engineering and Technology 48 2008, pp.190-197, DOI: 10.1.1.140.5102

[6] Tamanna Bhatia, "Link Analysis Algorithm for Web Mining," IJCST Vol. 2, Issue 2, June 2011, ISSN: 2319-5940

[7] Cooley, R, Mobasher, B., Srivastava, J., "Web Mining: Information and pattern discovery on the World Wide Web", IEEE 1997, pp.558-569, ISSN: 1082-3409

[8] Tsuyoshi, M and Saito, K., "Extracting User's Interest for Web Log Data", IEEE 2006, pp.343-346, ISBN: 0-7695-2747-7

[9] Fang Yuan, Li-Juan Wang, Ge Yu, "Study on Data Pre-processing Algorithm in Web Log Mining", IEEE Nov, 2003, pp.28-32 vol.1, ISBN: 0-7803-8131-9

[10] Fu, Y. and Shih, M., "A Framework for Personal Web Usage Mining", International Conference on Internet Computing, Las Vegas, NV, pp. 595-600, 2002, DOI: 10.1.1.90.8614