A MODEL OF DYNAMIC USER PATTERN CLUSTER USING TWO LEVEL PARTITIONING ALGORITHM

Ms. Rashmi Patel[1], Mrs. Risha Tiwari[2], Mr. Dushyantsinh Rathod[3]
[1]PG Student CE, [2]Asst. Prof., Computer Department, [3]Asst. Prof., CE/IT Department
[1,2]H.G.C.E., Vahelal, [3]ASOIT, Ahmedabad

Abstract: The expanded on-line applications are prompting to exponential development of the web content. The vast majority of the business associations are intrigued to know the web client conduct to improve their business. In this unique circumstance, clients route in static and element web applications assumes a vital part in comprehension client's interests.The static mining procedures may not be appropriate as it is for element web log documents and basic leadership. Conventional web log preprocessing approaches and weblog use designs have confinements to break down the substance association with the perusing history This thing, concentrates on different static web log preprocessing and mining strategies and their material confinements for element web mining using this techniques we can create pattern cluster so we can easily retrieve data from data source.In this paper I have just implemented 1st level algorithm only And in future work i have to create pattern cluster for dataset using 2nd level clustering. This algorithm increase 6% of performance, efficiency and accuracy.

Index Terms: Data Mining, Web log mining, Web Mining,Clustering, Pattern Clustering.

I. INTRODUCTION

Web mining is a process to analyze the online Web data, navigate between various Web sites and perform transaction of data across the Web. According to the types of data can be mined, web mining is classified into three types.Web Content Miningdiscovers information or knowledge from millions of sources across the Web. Web structure mining is the technique of finding structure information from the web. 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.

II. TYPES OF WEB LOG FILE FORMAT

- W3C(World Wide Web Consortium) Extended Log file Format Extended log is a customizable ASCII format which has different types of fields.
- MicrosoftIIS(Internet Information Services) Log fileFormat can record more information than the NCSA format.
- NCSA(National Centre for Supercomputing Application) Ordinary Log file Format which is available for Web sites but not for FTP sites.

III. EXISTING ARCHITECTURE

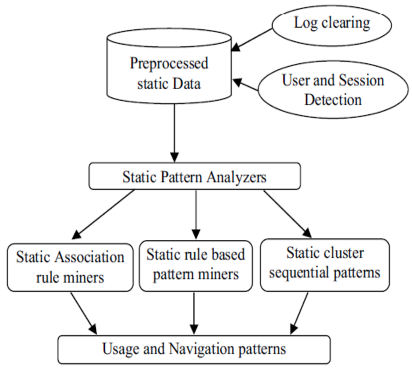

Log clearing
Preprocessed static Data
User and Session Detection
Static Pattern Analyzers
Static Association rule miners
Static rule based pattern miners
Static cluster sequential patterns
Usage and Navigation patterns

Figure.1:Existing Architecture

IV. EXISTING ALGORITHM

- Read N no of records from clean data source DS
- For i=1 to i<=N
- For each records R find frequent access data item F from data source DS
- Read frequency user access item F
- If R = F frequent records then
- Save for clustering frequent user access records in frequency access data source FDS
- Make cluster from frequency user access records
- Else not select records
- End If
- Next record

V. DRAWBACKS OF EXISTING

1. It does not provide clustering
2. Does not Cache of visited item
3. It recommended all visited item
4. Doesn’t create pattern clustering
5. It gives less performance
6. It consumes time
7. Low efficiency and less accuracy

VI. PROBLEM STATEMENT

Discuss the problem relating to Data cleaning of web log. Web log is generally noisy and ambiguous Web applications are increasing at an enormous speed and its users are increasing at exponential speed. Difficult to find the “right” or “interesting” information, There are a lot of work on data cleaning of web server logs irrelevant items and useless data can not completely removed. Difficulty in specifying the valid data from the log file with unlimited accesses to websites, web requests from multiple clients to multiple web servers.

VII. PROPOSED METHODOLOGY

The Two-level clustering method is improving the quality of data.

- The onelevel clustering is done in the form of datafrequently user access using clustering method. Remove unwanted or noisy data like .jpg, 404 page not found and any office file.
- The two level clustering is done by first changing the form of web log data into user access behavior patterns.

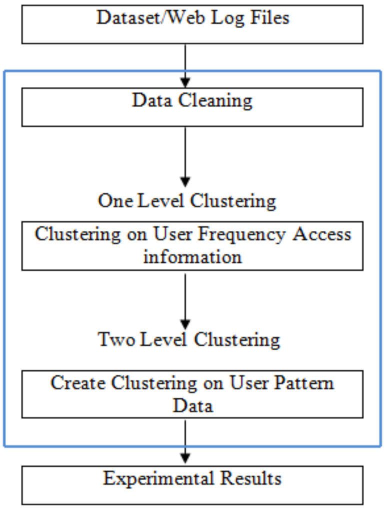


Figure.2:Proposed Clustering Process

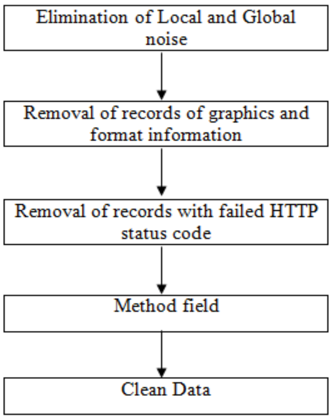

Elimination of Local and Global noise
Removal of records of graphics and format information
Removal of records with failed HTTP status code
Method field
Clean Data

Figure.3: Data Cleaning Steps

VIII. PROPOSED ALGORITHM

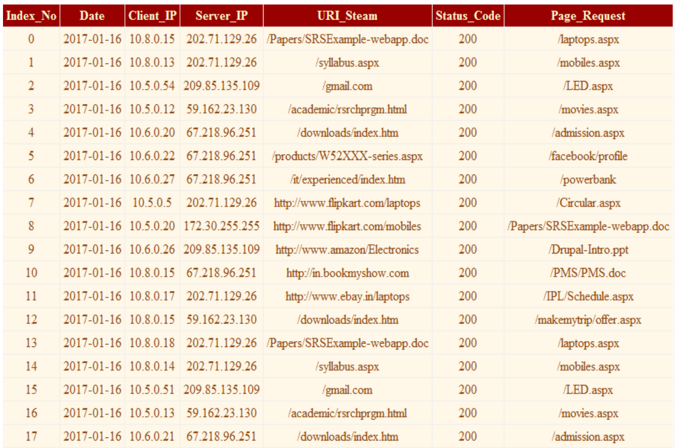
One Level Algorithm

```
1.  Read N no of  records from clean data source DS
    For i = 1 to i <= N
    Next
2.  For each records R find frequent access data item F
    from data source DS
3.  Remove unwanted or noisy data like .jpg, 404 page
    not found and any office file.
4.  Read frequency user access item F
5.  If R = F frequent records then
6.  Save for clustering frequent user access records in
    frequency access data source FDS
7.  Make cluster from frequency user access records
8.  Else not select records
9.  End If
10. Next record
```

Two Level Algorithm

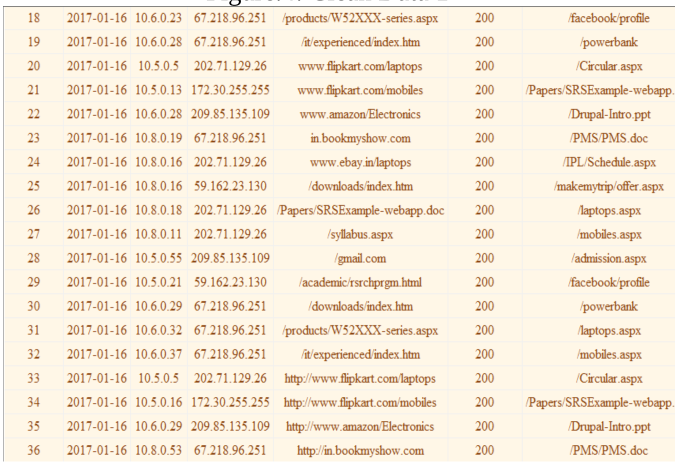
```
1.  Read N no of  records from clean data source FDS
    For i = 1 to i <= N
    Next
2.  For each records R from data source FDS find
    pattern data.
3.   Read pattern data using specified address from data
    source FDS.
4.   If requested records from frequent data source FDS
    with specified pattern then
5.  Collect and Save in pattern data source PDS.
6.  Make two level cluster in pattern data source PDS.
7.  Else not select that records.
8.  End If
9.  Next record
```

IX. OUTPUT



Index_No	Date	Client_IP	Server_IP	URI_Steam	Status_Code	Page_Request
0	2017-01-16	10.8.0.15	202.71.129.26	/Papers/SRSExample-webapp.doc	200	/laptops.aspx
1	2017-01-16	10.8.0.13	202.71.129.26	/syllabus.aspx	200	/mobiles.aspx
2	2017-01-16	10.5.0.54	209.85.135.109	/gmail.com	200	/LED.aspx
3	2017-01-16	10.5.0.12	59.162.23.130	/academic/rsrchprgm.html	200	/movies.aspx
4	2017-01-16	10.6.0.20	67.218.96.251	/downloads/index.htm	200	/admission.aspx
5	2017-01-16	10.6.0.22	67.218.96.251	/products/W52XXX-series.aspx	200	/facebook/profile
6	2017-01-16	10.6.0.27	67.218.96.251	/it/experienced/index.htm	200	/powerbank
7	2017-01-16	10.5.0.5	202.71.129.26	http://www.flipkart.com/laptops	200	/Circular.aspx
8	2017-01-16	10.5.0.20	172.30.255.255	http://www.flipkart.com/mobiles	200	/Papers/SRSExample-webapp.doc
9	2017-01-16	10.6.0.26	209.85.135.109	http://www.amazon/Electronics	200	/Drupal-Intro.ppt
10	2017-01-16	10.8.0.15	67.218.96.251	http://in.bookmyshow.com	200	/PMS/PMS.doc
11	2017-01-16	10.8.0.17	202.71.129.26	http://www.ebay.in/laptops	200	/IPL/Schedule.aspx
12	2017-01-16	10.8.0.15	59.162.23.130	/downloads/index.htm	200	/makemytrip/offer.aspx
13	2017-01-16	10.8.0.18	202.71.129.26	/Papers/SRSExample-webapp.doc	200	/laptops.aspx
14	2017-01-16	10.8.0.14	202.71.129.26	/syllabus.aspx	200	/mobiles.aspx
15	2017-01-16	10.5.0.51	209.85.135.109	/gmail.com	200	/LED.aspx
16	2017-01-16	10.5.0.13	59.162.23.130	/academic/rsrchprgm.html	200	/movies.aspx
17	2017-01-16	10.6.0.21	67.218.96.251	/downloads/index.htm	200	/admission.aspx

Figure.4: Clean Data 1



18	2017-01-16	10.6.0.23	67.218.96.251	/products/W52XXX-series.aspx	200	/facebook/profile
19	2017-01-16	10.6.0.28	67.218.96.251	/it/experienced/index.htm	200	/powerbank
20	2017-01-16	10.5.0.5	202.71.129.26	www.flipkart.com/laptops	200	/Circular.aspx
21	2017-01-16	10.5.0.13	172.30.255.255	www.flipkart.com/mobiles	200	/Papers/SRSExample-webapp.
22	2017-01-16	10.6.0.28	209.85.135.109	www.amazon/Electronics	200	/Drupal-Intro.ppt
23	2017-01-16	10.8.0.19	67.218.96.251	in.bookmyshow.com	200	/PMS/PMS.doc
24	2017-01-16	10.8.0.16	202.71.129.26	www.ebay.in/laptops	200	/IPL/Schedule.aspx
25	2017-01-16	10.8.0.16	59.162.23.130	/downloads/index.htm	200	/makemytrip/offer.aspx
26	2017-01-16	10.8.0.18	202.71.129.26	/Papers/SRSExample-webapp.doc	200	/laptops.aspx
27	2017-01-16	10.8.0.11	202.71.129.26	/syllabus.aspx	200	/mobiles.aspx
28	2017-01-16	10.5.0.55	209.85.135.109	/gmail.com	200	/admission.aspx
29	2017-01-16	10.5.0.21	59.162.23.130	/academic/rsrchprgm.html	200	/facebook/profile
30	2017-01-16	10.6.0.29	67.218.96.251	/downloads/index.htm	200	/powerbank
31	2017-01-16	10.6.0.32	67.218.96.251	/products/W52XXX-series.aspx	200	/laptops.aspx
32	2017-01-16	10.6.0.37	67.218.96.251	/it/experienced/index.htm	200	/mobiles.aspx
33	2017-01-16	10.5.0.5	202.71.129.26	http://www.flipkart.com/laptops	200	/Circular.aspx
34	2017-01-16	10.5.0.16	172.30.255.255	http://www.flipkart.com/mobiles	200	/Papers/SRSExample-webapp.
35	2017-01-16	10.6.0.29	209.85.135.109	http://www.amazon/Electronics	200	/Drupal-Intro.ppt
36	2017-01-16	10.8.0.53	67.218.96.251	http://in.bookmyshow.com	200	/PMS/PMS.doc

Figure.5:Clean Data 2

X. CONCLUSION AND FUTURE WORK

Data filtering perform by removing unwanted patterns from each record in the database. Since the pre-processing techniques performed is to mine the interesting patterns, the data end with *.jpg, *.gif, *.bmp be removed. In this paper I have just implemented 1st level algorithm only And in future work i have to create pattern cluster for dataset using 2nd level clustering. This algorithm increase 6% of performance, efficiency and accuracy.So we should format all this log files so we can easily make customizable combined file for analysis.

XI. REFERENCES

[1] P. Dhanalakshmi, Dr. K. Ramani, Dr. B. Eswara Reddy, “The Research of Preprocessing and Pattern Discovery Techniques on Web Log files” -978-1-4673-8286-1/16 $31.00, 2016 IEEE DOI 10.1109/IACC.2016.35

[2] ShashiMehrotra, ShrutiKohli, “Comparative Analysis of K-Means with other Clustering Algorithms to Improve Search Result” - 978-1-4673-7910-6/15/$31.00, 2015 IEEE

[3] Dilip Singh Sisodia, ShrishVerma, “Web Usage Pattern Analysis Through Web Logs: A Review” - 978-1-4673-1921-8/12/$31.00, 2012 IEEE

[4] TingZhong Wang, D. Jin and S. Lin (Eds.), “The Development of Web Log Mining Based on ImproveK-Means Clustering Analysis” - Advances in CSIE, Vol. 2, AISC 169, pp. 613–618. Springer-Verlag Berlin Heidelberg 2012

[5] S.C. Satapathy et al. (Eds.), “Design and Implementation of an Effective Web Server Log Preprocessing System” -Proceedings of the InConINDIA 2012, AISC 132, pp. 897–905. Springer-Verlag Berlin Heidelberg 2012

[6] J. Monisha Privthy Jeba, M. S. Bhuvaneswari, K.Muneeswaran, “Extracting Usage Patterns from Web Server” - 978-1-4673-6615-1/16/$31.00 © 2016 IEEE

[7] Supinder Singh, SukhpreetKaur, “Web Log File Data Clustering Using K-Means and Decision Tree” - 2013, IJARCSSE All Rights Reserved

[8] Ripal Patel, Mr. KrunalPanchal, Mr. DushyantsinhRathod, “Efficient Log Mining from Web Server Using Clustering Technique” - JJETIR1512016 - 2015

[9] Mr. Dushyant B. Rathod, “Customizable Web Log Mining from Web Server Log” - IJEDR1302021

[10] CiroGracia, Xavier Anguera, Xavier Binefa, “Two-level clustering towards unsupervised discovery of acoustic classes ” – IEEE - 2013

[11] Mr. DushyantsinhRathod, Dr. SamratKhanna, Mr. VijaykumarGadhavi “A Survey on Different Efficient Clustering Techniques used in Web” ISJRD - 2016

[12] http://www.w3.org/TR/WD-logfile.html

[13] http://www.extratrend.com