Improved Two Level K-Means Clustering Algorithm to Generate User Pattern Clustering

Mr. Dushyantsinh Rathod[1], Dr. Samrat Khanna[2]
[1] Rai University,Dholka
[2] ISTAR,VVNAGAR

Abstract- *Data cleaning perform in the Data Preprocessing and Web Usage Mining. The work on data cleaning of web server logs irrelevant items and useless data can not completely removed and Overlapped data causes difficulty during page ranking. Previous paper had given 30% performance of web log data. So we have Implemented Two-level clustering method to get pattern data for mining. This paper presents WebLogCleaner can filter out much irrelevant, inconsistent data based on the common of their URLs and it is going to improving 60% of the data quality, performance and efficiency of Web Log files.*

Keywords- Web Usage Mining (WUM); Data cleaning; web log mining ; Web Page Mining; Preprocessing.

I. INTRODUCTION

Data mining is the computational process of discovering patterns in large amount data sets involving methods at the intersection of artificial intelligence, machine learning of Data System. The World Wide Web is now a huge database with this growth there arises a need for analyzing the data. The process of discovery and analysis of Web is called Web mining. Web mining is the application of data mining techniques to discover patterns from the Web. Web mining can be divided into three different types 1) Web Structure Mining 2) Web Content Mining 3) Web Usage Mining. Web structure mining is the process of discovering the connection between web pages. Web content mining includes mining, extraction and integration of useful data and knowledge of Web page content. Web Usage Mining is a technique of extracting useful information from the Web Log, e.g. the pattern in which a user goes through different Webpages. WebLogCleaner that can filter out plenty of irrelevant items based on the common prefixof their URLs of data cleaning methods. Mining enterprise proxy log plays an important role for enterprise manager and employer whichmakes it difficult to find the “right” or “interesting” information [1]. Web Log are generally noisy and ambiguous. Web applications are increasing at an enormous speed and its users, are increasing at exponential speed.

There are lots of work on data cleaning of web server logs irrelevant items and useless data can not completely removed. When multiple data sources need to be integrated, data quality problems are present in single data collections, such as files and databases. WebLogCleaner that can filter out plenty of irrelevant items based on the common prefix of their URLs. This method is improving data quality by removing the irrelevant items. It is described of data characteristics reveals the importance and difficulty of data cleaning in web mining.

II. WEB USAGE MINING

Web Usage Mining could be a technique of extracting useful information from the web log, e.g. the pattern in which a user goes through different Web Pages. Using usage mining a designer can work on improving the web site or to provide a personalized service. Web Usage Mining consists of three steps [6].

1) Data Preprocessing
2) Pattern Discovery
3) Generate Cluster Pattern

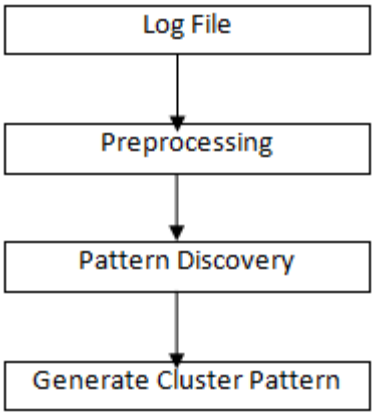


Figure.1 WebLog Mining Process

1. Data Preprocessing

The preprocessing of web logs is complex and time consuming and it is done using the following steps. The main task of data preprocessing is to select standardized data from the original log files, prepared for user navigation pattern discovery algorithm [5].

1) Data Cleaning
2) Page view Identification
3) Path Completion

4) Formatting

1.1 Data Cleaning

Data cleansing is that the method of removing irrelevant logs from log entries. Since HTTP is a connectionless protocol, when a user browse a web page in several log entire graphics and scripts are downloaded along with the HTML file. Data cleaning involves:-

a) Removal of Global and local Noise
b) Removal of images, video etc.
c) Removal of records that failed HTTP status code
d) Robots cleaning
e) Web noise can be normally categorized into two groups depending on their granularities.
f) Global Noise are corresponds to the unnecessary objects with huge granularities, which are no smaller than individual pages.
g) Local (Intra-Page) Noise are corresponds to the irrelevant items inside a Web page.

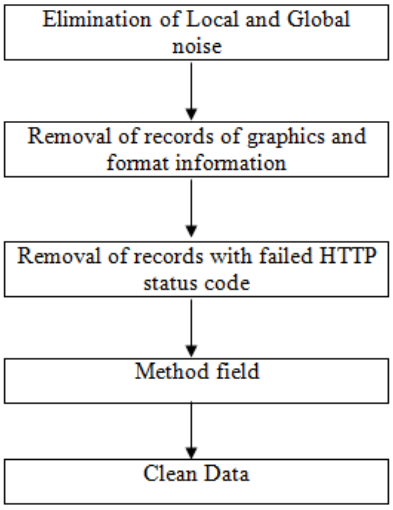


Figure 2 - Data cleansing steps

1.2 Page view Identification

Page view is a collection of web object. Page view identification is the process of identifying which page access files belong to a single page view. All the page views are assigned with a page view id.

2. Pattern Discovery

It is a method used in various fields such as data mining, pattern recognition, etc. pattern discovery involves finding a pattern in which the web user uses the web. There are various algorithms available to do this process such as the Association Rule for data mining.

3. Pattern Analysis. It is the last step in mining. It involves analyzing the pattern that is discovered in pattern discovery process. Useful and interesting pattern are kept and rest of the pattern, which are least useful, and interesting are removed.

III. 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.

Overlapped data cause difficulty during Page Ranking, When multiple data sources need to be integrated, data quality issues are present in single data collections, like files and databases, e.g., because of misspellings during data entry, missing information or alternative invalid data. The Standard Log filecontains irrelevant inconsistent data. Difficulty of knowledge extraction during Web Log Mining.

IV. TWO-LEVEL K-MEANS CLUSTERING

In this paper I Implemented Improved method of clustering, which is used to generate cluster pattern data.

Two-level K-means clustering method

The Two-level clustering method is improving the quality of data in the WUM process, which is the two-level clustering. Based on the results of two level clustering method on web log data, it can be concluded that this method can improve the quality of data web log.

- The first level clustering is done in the form of data frequently user access using non-hierarchical clustering method.

- The second level clustering is done by first changing the form of web log data into user access behavior patterns.

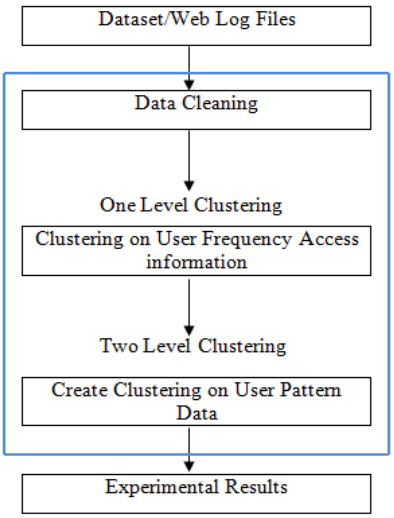


Figure 3: Two level clustering process

V. IMPROVED ALGORITHMS

1) ONE LEVEL CLUSTERING 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 item ***F*** from data source ***DS***
3. Read frequency user access items ***F.***
4. If ***R=F*** frequent records then
5. Save for clustering frequent user access records in frequency access data source ***FDS***
6. Make one level cluster from frequency user access records
7. Else not select records
8. End if
9. Next record

2) TWO LEVEL CLUSTERING 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

VI. RESULTS

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/rsrchprgm.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/W52XXX-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/rsrchprgm.html	200	/movies.aspx
17	2015-08-13	10.6.0.21	67.218.96.251	/downloads/index.htm	200	/admission.aspx

Fig 1. Final Clean Data

Index_No	Date	Client_IP	Server_IP	URI_Steam	Status_Code	Page_Request	Flag
0	2015-08-13	10.8.0.15	202.71.129.26	/Papers/SRSExample-webapp.doc	404	/samsung.jpg	1
1	2015-08-13	10.8.0.13	202.71.129.26	/syllabus.aspx	404	/LG.jpg	1
2	2015-08-13	10.5.0.54	209.85.135.109	/gmail.com	404	/LED.aspx	1
3	2015-08-13	10.5.0.12	59.162.23.130	/academic/rsrchprgm.html	404	/samsung.jpg	1
4	2015-08-13	10.6.0.20	67.218.96.251	/downloads/index.htm	404	/admission.aspx	1
5	2015-08-13	10.6.0.22	67.218.96.251	/products/W52XXX-series.aspx	404	/facebook/profile	1
6	2015-08-13	10.6.0.27	67.218.96.251	/it/experienced/index.htm	404	/powerbank	1
7	08/13/2015	10.5.0.5	202.71.129.26	http://www.flipkart.com/laptops	404	/Circular.aspx	1
8	08/13/2015	10.5.0.20	172.30.255.255	http://www.flipkart.com/mobiles	404	/Papers/SRSExample-webapp.doc	1
9	08/13/2015	10.6.0.26	209.85.135.109	http://www.amazon/Electronics	404	/Drupal-Intro.ppt	1
10	08/13/2015	10.8.0.15	67.218.96.251	http://in.bookmyshow.com	404	/PMS/PMS.doc	1
11	08/13/2015	10.8.0.17	202.71.129.26	http://www.ebay.in/laptops	404	/IPL/Schedule.aspx	1
12	08/13/2015	10.8.0.15	59.162.23.130	/downloads/index.htm	404	/makemytrip/offer.aspx	1
13	2015-08-13	10.8.0.18	202.71.129.26	/Papers/SRSExample-webapp.doc	404	/laptops.aspx	1
14	2015-08-13	10.8.0.14	202.71.129.26	/syllabus.aspx	404	/mobiles.aspx	1
15	2015-08-13	10.5.0.51	209.85.135.109	/gmail.com	404	/LED.aspx	1
16	2015-08-13	10.5.0.13	59.162.23.130	/academic/rsrchprgm.html	404	/movies.aspx	1
17	2015-08-13	10.6.0.21	67.218.96.251	/downloads/index.htm	404	/admission.aspx	1

Fig.2 Noisy Data with Flag Storage

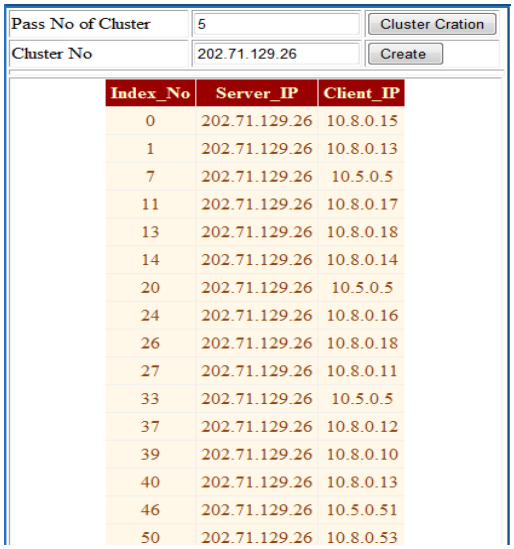


Fig.3 (Pattern Cluster 1)

Pass No of Cluster 5 Cluster Cration

Cluster No 209.85.135.109 Create

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.4 (Pattern Cluster 2)

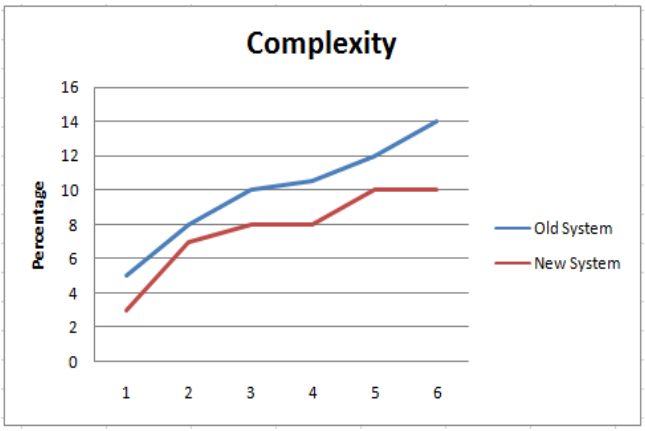
Pass No of Cluster 5 Cluster Cration

Cluster No 67.218.96.251 Create

Index_No	Server_IP	Client_IP
4	67.218.96.251	10.6.0.20
5	67.218.96.251	10.6.0.22
6	67.218.96.251	10.6.0.27
10	67.218.96.251	10.8.0.15
17	67.218.96.251	10.6.0.21
18	67.218.96.251	10.6.0.23
19	67.218.96.251	10.6.0.28
23	67.218.96.251	10.8.0.19
30	67.218.96.251	10.6.0.29
31	67.218.96.251	10.6.0.32
32	67.218.96.251	10.6.0.37
36	67.218.96.251	10.8.0.53
43	67.218.96.251	10.6.0.41
44	67.218.96.251	10.6.0.42
45	67.218.96.251	10.6.0.47
49	67.218.96.251	10.8.0.55

Fig.5 (Pattern Cluster 4)

VII. COMPARISON CHART



Above performance charts shows that the performance is increasing 60% Quality of data as compared to previous algorithm as 30% quality of data and complexity charts shows that when performance is increasing then by default complexity is decreasing.

VIII. CONCLUSION AND FUTURE WORK

There are many techniques proposed by totally different researchers for the web usage mining. This paper mentioned about Two-level clustering method available for web usage mining.

This previous paper has attempted to give EPFLog Miner quality is about 30% Performance of weblog mining. Where these new algorithms gives 60% performance of WebLogMiner and decreasing the complexity of web log data. Web log mining consists of data preprocessing, pattern discovery and Cluster generation. The results of Web Log mining can be used for various applications such as web personalization, site recommendation, site improvement, etc.
In this paper, I describe Two-level Clustering Algorithm for web log preprocessing techniques. In the future work apply this algorithm on Personalize Web recommended system to get accuracy and efficiency based on different criteria using pattern mining

REFERENCES

[1] HongzhouShaa,c, TingwenLiub,c, Peng Qinb,c Yong Sunb,c, QingyunLiub,c, “WebLogCleaner: Improving Data Quality of Enterprise Proxy Logs for Efficient Web Usage Mining”, Information Technology and Quantitative Management , ITQM 2013 Procedia Computer Science 17 (2013).

[2] T. Hussain, S. Asghar, N. Masood, “Web Usage Mining: A Survey on Preprocessing of Web Log File”, in:

Proceedings of the 2010 InternationalConference on Information and Emerging Technologies (ICIET), IEEE, 2010, pp. 1–6.

[3] N. Tyagi, A. Solanki, S. Tyagi, "An Algorithmic Approach to Data Preprocessing in Web Usage Mining", International Journal of Information Technology and Knowledge Management 2 (2) (2010) 279–283.

[4] Ling Zheng Hui Gui. Feng Li, "Optimized Data Preprocessing Technology for Web Log Mining", InternationalConference On Computer Design And Applications (ICCDA 2010).

[5] Michal Munk, JozefKapustaa, Peter Šveca*, "Data Preprocessing Evaluation for Web Log Mining: Reconstruction of Activities of a Web Visitor" International Conference on Computational Science, ICCS 2011 Procedia Computer Science 1 (2012).

[6] P.Nithya, Dr.P.Sumathi , "Novel Pre-Processing Technique for Web Log Mining by Removing Global Noise and Web Robots. " National Conference on Computing and Communication Systems (NCCCS) IEEE 2012.

[7] V. SUJATHAa, PUNITHAVALLIb, a*, "Improved user navigation pattern Prediction technique from web log data."International Conference on Communication Technology and System Design 2011 Procedia Engineering 30 (2012) 92.

[8] Theint Theint Aye, "Web Log Cleaning for Mining of Web Usage Patterns."IEEE 2011.

[9] Chu-Hui Lee, Yu-lung Lo, Yu-Hsiang Fu, "A novel prediction model based on hierarchical characteristic of web site", Elsevier (2010), Expert Systems with Applications 38 (2011) 3422–3430.

[10] Vijayashri Losarwar, Dr. Madhuri Joshi, "Data Preprocessing in Web Usage Mining", International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July 15-16, 2012.

[11] RohitAgarwal,K. V. Arya, Shashi Shekhar, Rakesh Kumar, "An Efficient Weighted Algorithm for Web Information Retrieval System", IEEE(2011).

[12] Tasawar Hussain, Dr. SohailAsghar, Dr. Nayyer Masood, "Web Usage Mining: A Survey on Preprocessing of Web Log File", IEEE (2010).

[13] 1Yuhefizar,2Budi Santosa,3I Ketut Eddy P.,4Yoon K. Suprapto, "Two Level Clustering Approach for Data Quality Improvement in Web Usage Mining", Journal of Theoretical and Applied Information Technology 20th April 2014.

[14] 1B.Uma Maheswari, 2 Dr. P.Sumathi, "A New Clustering and Preprocessing for Web Log Mining", IEEE (2014).

[15] Rana Forsati, Mohammad Reza Meybodi,AfsanehRahbar, "An Efficient Algorithm for Web Recommendation Systems", IEEE (2009)