

A Comparison of K-Means Clustering and Smart Two Level K-Means Clustering Algorithm

Mr. Dushyantsinh Rathod¹, Dr. Samrat Khanna², Mr. Manish Singh³

^{1,2,3}Department of CE/IT

^{1,2}Rai University

³ASOIT

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. K-means clustering algorithm had given 30% performance of web log data and Two-level clustering method to get pattern data for mining. It is going to improving 8% 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

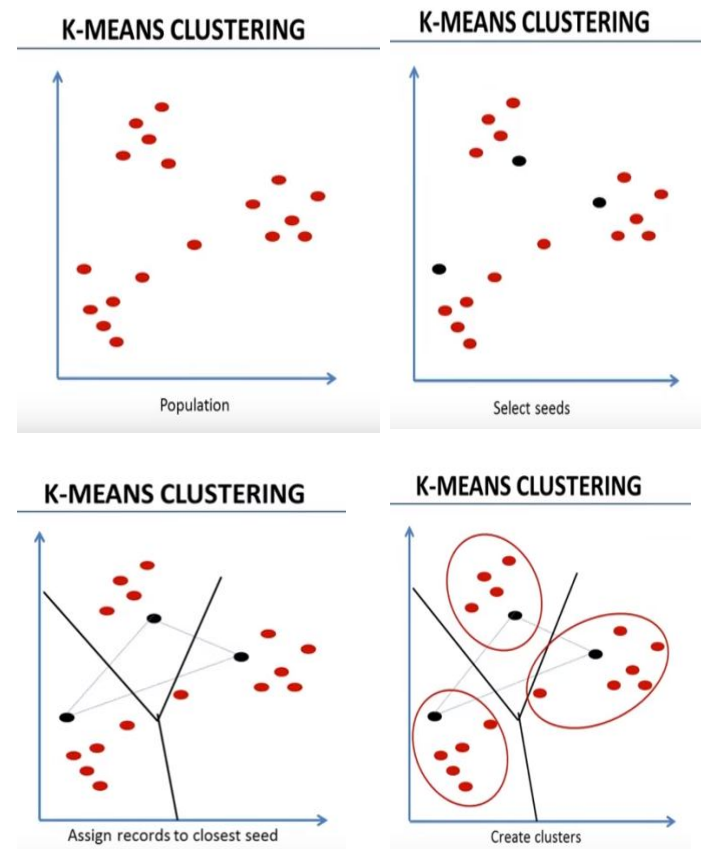
Data mining is the computational process of discovering patterns in large amount data sets involving methods at the intersection of artificial intelligence, machine learning of Data System. The World Wide Web is now a huge database with this growth there arises a need for analyzing the data. The process 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. Weblog Cleaner that can filter out plenty of irrelevant items based on the common prefix of their URLs of data cleaning methods. Mining enterprise proxy log plays an important role for enterprise manager and employer which makes 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. Weblog Cleaner 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 reveal the importance and difficulty of data cleaning in web mining.

II. K-MEANS CLUSTERING

K-means clustering is just making the clustering based on clean data. Fig. shows the how to make a cluster using seed points or attributes.



III. OLD ALGORITHM

1. Read records from cleaned data source
2. For each records find frequent access data from data source
3. Read frequency user access data.
4. If frequent records found then
5. Save for clustering frequent user access records in frequency access data source
6. Make one level cluster from frequency user access records
7. Else not select records
8. End if
9. Next record

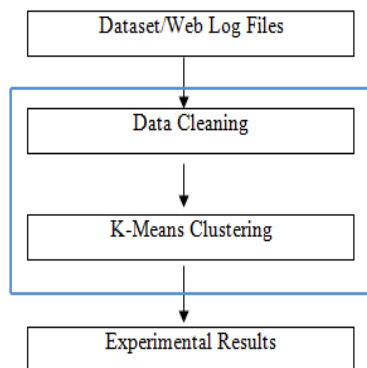


Fig.1.K-means clustering

IV. TWO-LEVEL K-MEANS CLUSTERING

A. 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.

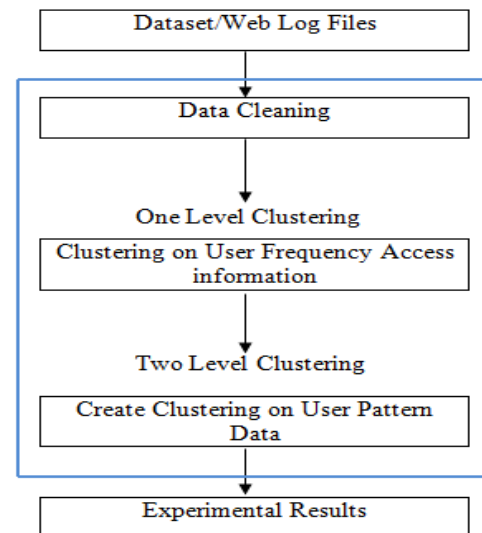


Fig. 3. Two Level clustering process

V. IMPROVED ALGORITHMS

A. One Level Clustering Algorithm

1. Read N no of records from clean data source DS
For $i= 1$ to $i \leq N$
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

B. Two Level Clustering Algorithm

1. Read N no of records from clean data source FDS
For $i= 1$ to $i \leq N$
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.RESULT

Index_No	Date	Client_IP	Server_IP	URI_Steam	Status_Code	Request
0	2015-08-13	10.8.0.15	202.71.129.26	/Papers/SRSEExample-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/SRSEExample-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.bookmystow.com	200	/PMS/PMS.doc
11	08/13/2015	10.8.0.17	202.71.129.26	http://www.ebay.in/laptops	200	/PL/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/SRSEExample-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 4. 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/SRSEExample-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/SRSEExample-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.bookmystow.com	404	/PMS/PMS.doc	1
11	08/13/2015	10.8.0.17	202.71.129.26	http://www.ebay.in/laptops	404	/PL/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/SRSEExample-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

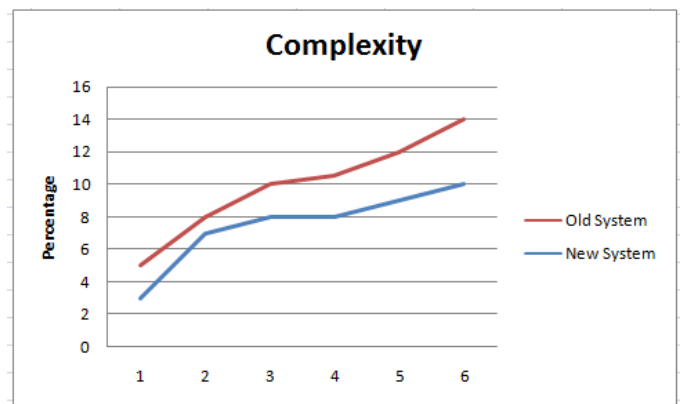
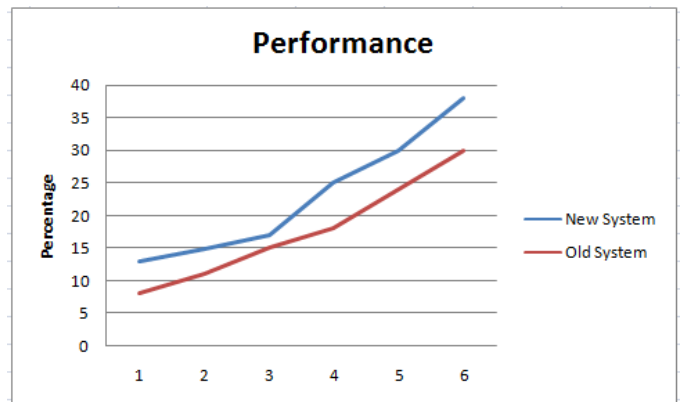
Fig.5. Noisy Data with Flag Storage

Pass No of Cluster	5	Cluster Cration
Cluster No	202.71.129.26	Create

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.6. (Pattern Cluster 2)

VII. COMPARISON CHART



Above performance charts shows that the performance is increasing 8% Quality of data from 38% 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.

VII. 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 8% performance of Weblog Miner from total 38% 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 (ICDDA 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] TheintTheint 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] VijayashriLosarwar, 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. Suprpto, “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)