MODIFIED K-MEDIOD ALGORITHM OF CLUSTERING FOR WEB USAGE MINING

Sushma Kharkar[1], Ankita Gandhi[2]

[1] *Student, Information Technology, Parul Institute of Engineering and Technology, Gujarat, India*
[2]*Assistant Professor, Computer Science and Technology, Parul Institute of Engineering and Technology, Gujarat, India*

ABSTRACT

World Wide Web is consisting of large amount of information and provides it different kind's users in server. Due to increasing number of users access the web day by day, there is a need to analyses behavior of such user, in order to monitoring and improving the performance and throughput of website. Web usage mining is one of the parts of data mining applications which dealing with log files and extracting the useful information from web data. There are different phases included in web usage mining: Data pre-processing, discover pattern and pattern analysis. Among all of them data pre-processing is the most crucial and time consuming phase of web usage mining because without good quality of data it is difficult to identify pattern of user's behavior. Pre-processing technique improves the quality and accuracy of the pattern mining algorithms. The existing system have done the pre-processing activities for reducing the size of the log file and to identify the unique users and sessions but existing system face scalability problem because web data increase day by day. The propose technique reduces the scalability problem and also improve the performance of web usage mining and pre-processing by using modified k-mediod clustering technique.

Keyword: - *Preprocessing, Data cleaning, Modified K-mediod, Web usage mining*

1. INTRODUCTION

In today's world websites remain the main source of information in every-day life. Thus there is a huge development of World Wide Web in its capacity of traffic and the size and complexity of web sites. Web mining is the application of data mining and artificial intelligence and so on to the web data identifies user's visiting behaviors and extracts their interests by means of patterns. Due to its normal application in Web analytics, e-learning, e-commerce and information retrieval. Web mining has been one of the significant areas in computer and information science. The application of Web Usage Mining systems in log data is to abstract the comportment of users which is used in variation of applications like pre-fetching, creation of attractive web sites, personalization, adaptive web sites, customer profiling, etc. Web mining is a data mining techniques to automatically discover and extract information from Web documents and services. There are three general classes of information that can be discovered by web mining: Web activity, from server logs and Web browser activity tracking. Web mining is used to understand customer behavior, evaluate the effectiveness of a particular Web site, and help quantify the success of a marketing campaign. According to mining objects there are three main types of web mining.

1. Web Content Mining
2. Web Structure Mining
3. Web Usage Mining

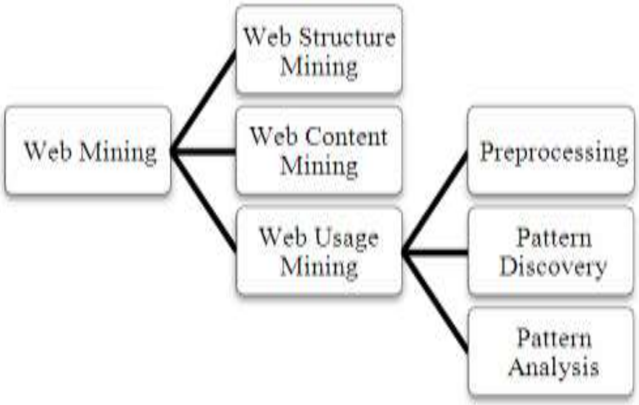


Figure- 1. Classification of Web Mining

1.1 web usage mining

Web usage mining is an important part of data mining Techniques which finds the interesting usage patterns from web data. Web usage mining is useful for discover user navigation patterns from web data; it also tries to discovers the useful information from the secondary data derived from the interactions with the users while searching on the Web. Web usage mining collects the data from Web log records to discover user access patterns of web pages.
Web usage mining includes three main steps:

1. Data Pre-processing
2. Pattern Discovery
3. Pattern Analysis

2. OVERVIEW ON EXISTING SYSTEM

The data preprocessing is the primary step in the data preparation method, objectives to reformat the unique logs to identify user's sessions. This process is maximum time consuming and intensive phase. A user session file is an input to the web usage mining process that provides information on who read the page of a web site, what pages accessed, the order in which the pages accessed and entire time spent on each page. Web server marks information whenever a user requests a reserve from the site. A web server normally stores all user based activities of the web site in the form of server logs. The server log files acts as a main data sources in Web usage mining, which contain - access logs of the web server , application server logs. The significant task in the pre-processing phase is "field extraction". The log files having log entries which represent the particular click. The log entry includes of several fields which need to be isolated for additional processing.

Log data changes from other datasets used in data mining, and there are several difficulties which must be addressed in preparation for data mining. The core problem is to get a reliable dataset for mining process. Thus the data must be pre-treated and users accessing behaviour is to be created as transactions. These transactions are to be consistent. The Common log formats (CLF) or Extended Log Formats (ELF) only accounts the visitors browsing activities rather than the particulars of the visitor's uniqueness. This means that dissimilar visitors sharing the similar host cannot be distinguished. If there are proxy servers, the problem became much severe. Users are identified simply by using Cookies and authentication mechanism. But users are not involved by these types of sites due to privacy and confidentiality.

The limitation of existing system is that it is restricted for few records and preprocessing is a time consuming steps it required an algorithm which is scalable where existing system lacks in scalability. Because of that we need better algorithm to improve the performance of the web usage mining

3. PROPOSED WORK

The proposed system focuses on data cleaning, session Identification process and building the transactions in preprocessing stage. In this research a clustering method is given for effectively constructing the reliable transactions in data preprocessing. This gives us the clustering process which uses modified k-mediod algorithm and improve the scalability problems in terms of execution time.

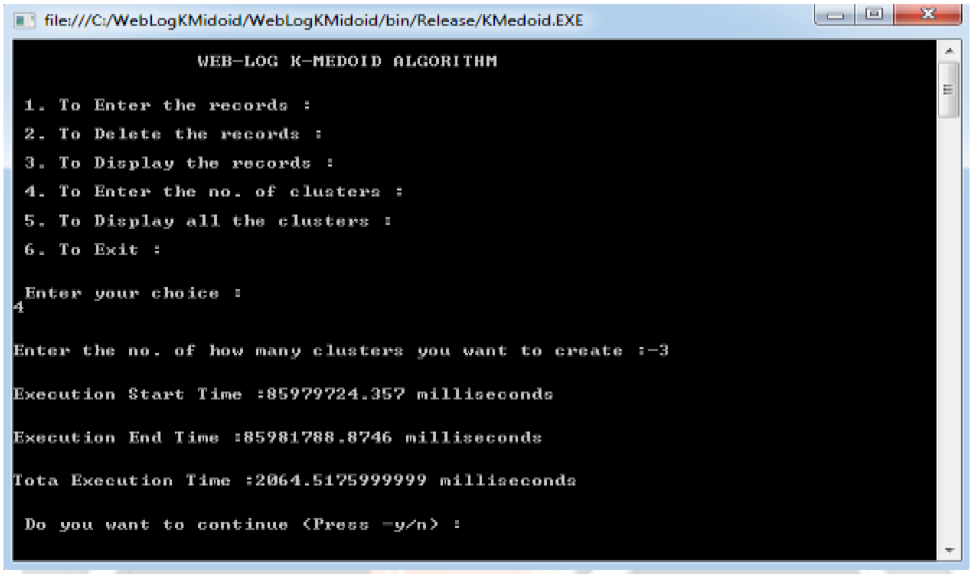
3.1 Clustering of data (Modified K-medoid)

Clustering of data means arranging the data in a similar type as a form database. All similar type of data in single cluster, different type of data in a different cluster in a database and user searching the data easily also user searching all time. K-medoid algorithm is used for clustering of common type of data element from n elements The *k*-medoids algorithm is a clustering Algorithm related to the *k*-means algorithm and the chooses data points as centers. Basically the *k*-means and *k*-medoids algorithms are partitions data into groups and both used to minimize the distance between points labeled in as cluster to a point as the center of that cluster. The algorithm as follows:

```
Input: k: The no.of clusters
D: A data set which containing n objects
Output: A set of k clusters that reduces the sum of the dissimilarities of all the objects to their adjacent medoid
1 Method: Arbitrarily choose k objects in D as the initial representative objects;
2 For each data-point di, find the closest centroid cj and assign di to cluster j Select initial
k=1;
3 Do
4 Old MSE=MSE;
5 MSE1=0;
6 For j=1 to k
7 mj=0; nj=0;
8 endfor
9 For i=1 to n
10 For j=1 to k
11 Compute squared Euclidean distance
11.1 Repeat allocates each remaining object to the cluster with the adjacent medoid;
11.2 If D < 0 then swap Oj with O arbitrary to the form the newest set of k medoid
12. Randomly select a non medoid object O arbitrary;
12.1 calculate the total points S of swapping object Oj with O ramdom;
12.2 Until no change
12.3 end for
13 Find the nearby centroid mj to xi;
14 mj=mj+xi; nj=nj+1;
15 MSE1= and is
16 endfor
17 For j=1 to k
18 nj=max (nj,1); mj=(mj/nj);
19 endfor
20 MSE=(MSE1); while (MSE<OldMSE)
```

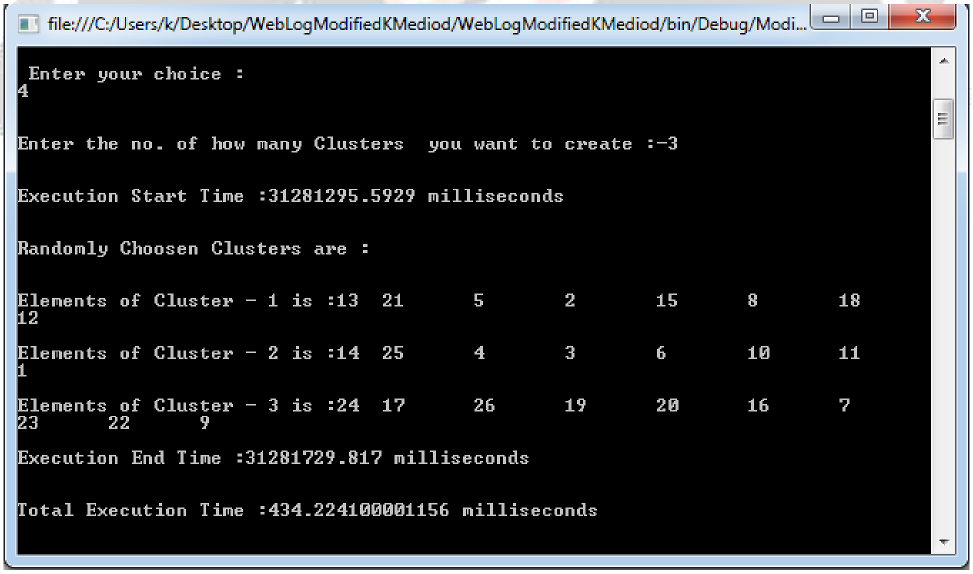
.

4. EXPERIMENTAL RESULTS


file:///C:/WebLogKMidoid/WebLogKMidoid/bin/Release/KMedoid.EXE

```
                WEB-LOG K-MEDOID ALGORITHM

 1. To Enter the records :
 2. To Delete the records :
 3. To Display the records :
 4. To Enter the no. of clusters :
 5. To Display all the clusters :
 6. To Exit :

 Enter your choice :
4

Enter the no. of how many clusters you want to create :-3

Execution Start Time :85979724.357 milliseconds

Execution End Time :85981788.8746 milliseconds

Tota Execution Time :2064.5175999999 milliseconds

 Do you want to continue (Press -y/n) :
```

Figure- 2: K-mediod Execution time for preprocessing


file:///C:/Users/k/Desktop/WebLogModifiedKMediod/WebLogModifiedKMediod/bin/Debug/Modi...

```
 Enter your choice :
4

Enter the no. of how many Clusters  you want to create :-3

Execution Start Time :31281295.5929 milliseconds

Randomly Choosen Clusters are :

Elements of Cluster - 1 is :13  21      5       2       15      8       18
12
Elements of Cluster - 2 is :14  25      4       3       6       10      11
1
Elements of Cluster - 3 is :24  17      26      19      20      16      7
23      22      9
Execution End Time :31281729.817 milliseconds

Total Execution Time :434.224100001156 milliseconds
```

Figure-3:Modified K-mediod Execution time for preprocessing

Table 5.3: Number of clusters and execution time for K-Mediod Algorithm and log Modified K-Mediod Algorithm

Number of clusters	Time taken to execute (In millisecond) Log K-Mediod Algorithm	Time taken to execute (In millisecond) Log Modified K-Mediod Algorithm
2	20322	16343
3	40381	26922
4	69234	48533
5	73345	49342

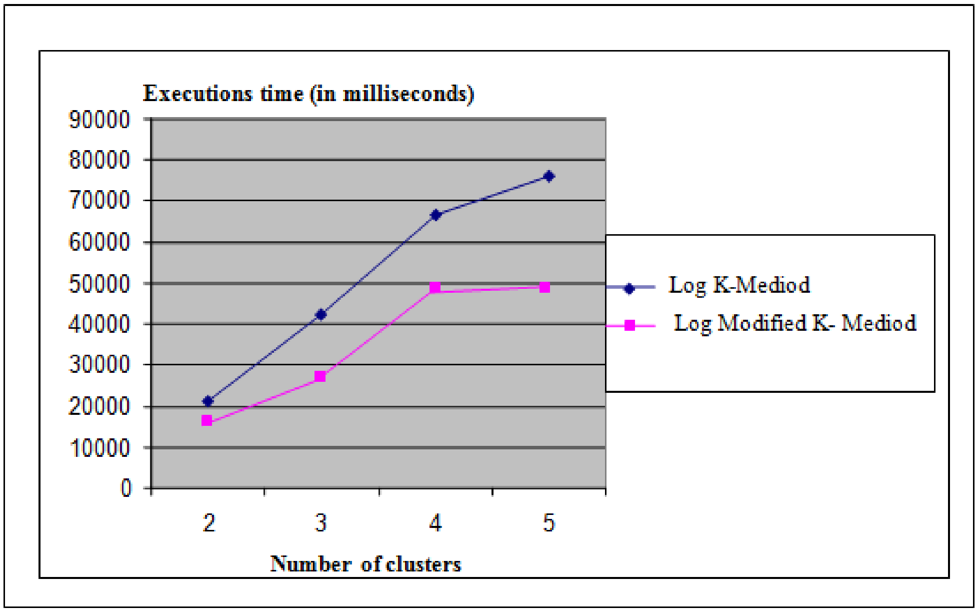


Figure 4- Graph Represent Number of Clusters and Execution Time for log K-Mediods Algorithm and log Modified K-Mediod Algorithm

Above Graph shows comparison between log K-mediod and log Modified K-Mediod Algorithm. As graph show that when number of clusters is less, log Modified K-Mediod Algorithm takes less time to execute than the log K-mediod. If the number of clusters is more than it is again true that log Modified K-Mediod Algorithm takes less time to execute than the K- mediod. At the particular number of the records the execution time taken by log Modified K-Mediod Algorithm takes approximately less time than log K- mediod

5. CONCLUSIONS

Preprocessing is important step of Web usage mining. In the present work, an attempt would be made to improve the quality of data of web log files. Preprocessing of web log file is first necessary and important process for web usage mining. Existing Algorithm is lacks in scalability problem. Usage data collection on the Web is incremental. Therefore, there is a need for mining algorithms to be scalable. Our propose work will improve the performance of preprocessing and increase the preprocessing speed and decreases execution time.

REFERENCES

[1] B.Uma Maheswari, Dr. P.Sumathi “A New Clustering and Preprocessing for Web Log Mining” 2014 IEEE World Congress on Computing and Communication Technologies

[2] Ying Han, Kejian Xia “Data preprocessing method based on user characteristic of interests For Web log mining” 2014 IEEE Fourth International Conference on Instrumentation and Measurement, Computer, Communication and Control

[3] Lya Hulliyyatus Suadaa “A Survey on Web Usage Mining Techniques and Applications” 2014IEEE International Conference on Information Technology Systems and Innovation

[4] Sheetal A. Raiyani, Shailendra Jain “Enhance Preprocessing Technique Distinct User Identification using Web Log Usage data” International Journal of Computer Science & Communication Networks, Vol 2.

[5] K. Sudheer Reddy, G. Partha Saradhi Varma, and M. Kantha Reddy “An Effective Preprocessing Method for Web Usage Mining” International Journal of Computer Theory and Engineering, Vol. 6, No. 5, October 2014

[6] Neha Goel, Dr. C.K.Jha “Preprocessing Web logs: A Critical phase in Web Usage Mining” IEEE 2015 International Conference on Advances in Computer Engineering and Applications.

[7] Harmit kaur, Hardeep singh “A Survey of Preprocessing Method for Web Usage Mining Process” 2014 International Journal of Computer Trends and Technology

[8] V.chitraa Dr.antony selvadoss thanamani “An Enhanced Clustering Technique for WebUsage Mining”2012 International Journal of Engineering Research & Technology(IJERT)

[9] Gurpreet Kaur “Accurate Analysis of Weblog Server File by Using Clustering Technique”2013 International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)

[10] Mr Jitendra B. Upadhyay, Dr. S. V. Patel “A Review Analysis of Preprocessing Techniques in Web usage Mining”2015 International Journal of Engineering Research & Technology (IJERT)

[11]P. Nithya, Dr. P. Sumathi “Novel Preprocessing Technique for Web Log Mining by removing Noise and Web Robots”2012 IEEE National Conference and Computing and Communication System.

[12]Zhang Huiying, Liang Wei “An Intelligent Algorithm of Data Preprocessing in Web Usage Mining”2004 IEEE.

[13] G.Shivprasad, N.V.Subha Reddy, U. Dinesh Acharya and Prakash K.Aithal “Neuro- Fuzzy Based Hybrid Model For Web Usage Mining”2015 ScienceDirect.

[14] Michal Munk.Jozwf Kapusta, Peter svec “Data Preprocessing for Web Log Mining Reconstruction of Activities of Web Visitors”2010 ScienceDirect.

[15] Nirali Honest and Dr.Atul Patel, Dr. Bankim Patel "A study of Path Completion in Web Usage Mining"2015 IEEE

[16] Natheer Khasawnch,Chien-Chung Chan "Achive User-based and Ontology-Based Web data Pre-processing for Web Usage Mining"2006 IEEE

[17] R. Lokeshkumar, P. Sengottuvelan "A Novel Approach to Improve Users Search Goal in Web Usage Mining"2015 International Scholarly and Scientific Research & Innovation

Websites:
1. https://en.wikipedia.org/wiki/Web_mining
2. https://en.wikipedia.org/wiki/Data_mining
3. http://webcache.googleusercontent.com/search?q=cache:http://dictionary.cambridge.org/dictionary/english/web-log-file&gws_rd=cr&ei=ieZoVrXLGsSSuATbw7WoDQ