IJMDRR
E- ISSN –2395-1885
ISSN -2395-1877

Research Paper
Impact Factor - 2.262
Peer Reviewed Journal

A COMPARATIVE STUDY ON K-MEANS AND K-MEDOID ALGORITHMS

Ms.R.Sangeetha
Assistant Professor of CS, K.S.R College of Arts & Science (Autonomous), Tiruchengode.

Abstract
Data mining is the process of "The nontrivial extraction of implicit, previously unknown, and potentially useful information from data". Web Mining is one of the concepts of data mining techniques. It is used extract knowledge from Web data, including web documents, hyperlinks between documents, usage logs of web sites, etc. When the user requests the record, the web server traces and accumulates the data regarding user interactions. In order to improve the website design, the analyziation of Web access logs of various web sites plays a major role. The variety of helpful information such as URL, IP address, time etc., can be collected by web log. Analyzing and discovering log help us to find more potential users of the web site and trace service quality of the site. Clustering is one of the concepts of data mining techniques. It is used in many research area fields. Clustering has been defined as "The objects can be classified into different groups based on partitioning sets of data into a subset". The clustering algorithm is divided into two methods such as Partition Method and Hierarchical Method. In this paper, we discuss about Partition Method and its algorithms. There are two types of partitioning algorithms that is is k-medoid and k-mean. It is a process of forming the disjoint clusters by grouping data objects therefore the data in each cluster are same, however different to the other clusters.

Keywords: Clustering, K-Mean, K-Medoid, Log Files.

I. INTRODUCTION

Data Mining is the process of extract of useful information from a huge volume of data. It is used to show the hidden knowledge of data and also implementing the real time applications. Data mining contain different algorithms for data analysis. Several data mining techniques are used for analysis the data such as Clustering, Association, and Classification etc. Compare with these clustering is one of the most efficient techniques for data analysis. The clustering method is subdivided into four: partitioning, hierarchical, grid-based and model-based methods. Partition based clustering generates a partition of the data such that objects in a cluster are more similar to each other than they are to objects in other clusters . Clustering techniques is used to extract hidden data from the exists data sets. It is the method of forming disjoint clusters by combining the data objects so that the individual cluster consists of similar data but different to the other clusters[1].

K-Means algorithm is used to minimizes the clustering error. It is one of the clustering method. It is easy and fast. So it is very attractive to the user. In this method, the input dataset is partitioned into K clusters. Each cluster is identified by centroid(also called cluster centre), It will be changing based on the data set. The initial values are named seed points. K-Means algorithm is used to calculate the distance between the inputs(data points) and centroid, and assign to the nearest centroid. Normally K-means algorithm has two problem:1. The number of cluster is unknown 2. Initial seed problem. It is very difficult to find out the web log files directly. To get useful information for the web logs, we need some other Some preprocessing techniques and pattern discovery algorithms. The data is split into groups within the direction of parameters. It is the best way for all the time. So the user can use different types of clustering algorithm to web log files to group them either link based, user based or session based.

Web Usage Mining is used to collect the details and activities of the website visitors. These type of research field for satisfying customer expectations. In the Common Log File format, the navigation details are maintained in web servers, proxy servers and client machines. The user's viewing website details are collected in various sources in CLF format. Maintaning the proper format for the web log is tedious due to its large size. So the following techniques/algorithms are used for different purpose.

1. Preprocessing-to extract knowledge for suitable weblog.
2. 2.Soft Clustering algorithms(fuzzy clustering)- accuracy and efficiency.
3. 3.K-Means clustering algorithm- Its simplicity, it is used to improve its efficiency .

Arbitrarily distributed input data points are used to evaluate the clustering quality and performance of two algorithms such as K-Means and K-Medoid. Because many of the researcher uses these two algorithms. The distance between two data points are taken to evaluate the clustering quality. In order to measure the performance of the algorithm, the computational time is calculated for each algorithm. K-Mediod algorithms provide better results compare with K-Means. It is demonstrate by the experimental results. The average execution time of the K-mediod algorithm is very less than the K-Means algorithm[4].

IJMDRR
E- ISSN –2395-1885
ISSN -2395-1877

To classifying cluster categories, K-Medoid algorithm is very efficient. K-Medoids clustering algorithm is very efficient in classifying cluster categories . Based on the algorithm analysis and chosen development of center point K, web model of ontology data set object is assigned in this paper.. This paper demonstrates through experiment results that the improved algorithm can enhance the accuracy of clustering results under semantic web. K-Medoids (partition clustering algorithm)which selects k clustering centers from data objects and set an initial partition nearest to clustering centre for other data before iterating and moving clustering centers continuously until an optimum partition is reached[5].

To classifying cluster categories, K-Medoid algorithm is very efficient. K-Medoids clustering algorithm is very efficient in classifying cluster categories . Based on the algorithm analysis and chosen development of center point K, web model of ontology data set object is assigned in this paper.. This paper demonstrates through experiment results that the improved algorithm can enhance the accuracy of clustering results under semantic web. K-Medoids (partition clustering algorithm)which selects k clustering centers from data objects and set an initial partition nearest to clustering centre for other data before iterating and moving clustering centers continuously until an optimum partition is reached[5].

K-Mediod clustering algorithm works like a K-means algorithm. Once the distance matrix is calculates by this algorithm then this distance matrix is used for finding new mediods at every step. K-Medoid algorithm show less time compare with K-Means algorithm[6].

The user session data are partition into set of clusters by K-Mediod algorithm act as a reduction mechanism. Similar scenario of user interactions represents by each cluster with the web application. Samples of each cluster collected and constructed for test data for web application test.

In this paper, the important issues of non-numeric data type of user sessions and their dissimilarity definition are addressed.

Various number of attributes are count in two client request, each one consists of basic request, none or many name-value pairs.It is accomplished by randomly selecting representative user sessions from each partition cluster without reconstructing or rearranging the user sessions data[7].

The analysis of K-Means and K-Mediod algorithms were examined based on their basic approach. The input data points are generated by two ways that is one by using normal distribution and another by applying uniform distribution. Randomly distributed data points were taken as input for these algorithms. For each category, the execution time of this algorithm is calculated then compared. The accuracy of the algorithm was determined during different execution of the program on the input data points. The average time of K-Medoid algorithm is less than the K-Means algorithm for both cases [8].

II. PARTITIONING TECHNIQUES

Among the various concepts, the partitioning technique concept is the one which divides the objects in multiple partitions. A single partition describes each cluster and a single cluster consists of objects having the similar characteristics and different cluster have objects having dissimilar characteristics. Both cluster types are used for dataset attributes. Partition technique has a unique feature of measuring distance and it is also used to identify similarity or dissimilarity of patterns between data objects[7]. K-Mean, K-Medoid and CLARAN are the three different algorithms of partition technique.

a. K-Mean

Centroid based technique is applied in K-Mean Algorithm. This technique accepts k input parameter and partition a set of n object from k clusters. Based upon the mean value of the object, the similarity among clusters is measured. The first step of algorithm is the random selection of k objects that represents cluster mean or centeroid.

Algorithm [10]: The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

- K:the number of clusters
- D:a data set containing n object

Output:

- A set of k clusters

IJMDRR
E- ISSN –2395-1885
ISSN -2395-1877

Method:
(a) Arbitrarily choose k objects from D as the initial cluster centers.
(b) Repeat
(c) Reassign each object to the cluster to which the object is the most similar,
Based on the mean value of the objects in the cluster;
(d) update the cluster means ,i.e., calculate the mean value of the objects for each cluster;
(e) Until no change;

b. K- Medoid

In order to represent the cluster the K-Mean method follows centroid techniques and it is sensitive to outliers which means the distribution of data may get disrupted due to a data object with an extremely large value. Therefore we utilized representative object technique based K-Medoid method to overcome the setback of disruption of data distribution by the data object with an extremely large value.

In this method to represent K cluster, K data objects are randomly selected as medoid and the leftover data objects are located in a cluster having nearest medoid to that data object. After completing the process of all data objects, new medoid is established that able to represent cluster in a better way and the entire process is repeated. Again based on the new medoids, all data objects are bound to the clusters. For each iteration, medoids modify their location step by step. Till the movement of any medoid, the process keeps going on. As a result, a set of n objects represents K clusters[3]. An algorithm for this method is given below.

Algorithm [10]: PAM, a k-medoids algorithm for partitioning based on medoid or central objects.

Input:
K: the number of clusters,
D: a data set containing n objects.

Output:
A set of k clusters.

Method:
(a) Arbitrarily choose k objects in D as the initial representative objects or seeds;
(b) Repeat
(c) Assign each remaining object to the cluster with the nearest representative object;
(d) Randomly select a non-representative object, Orandom.
(e) Compute the total cost of swapping representative object, Oj with Orandom;
(f) If S<0 then swap Oj with Orandom to form the new set of k representative object;
(g) Until no change;

III. EXPERIMENTAL RESULTS

Three data sets are used for our experimental results named D1, D2, D3. These data set consist of log files. Log files of clustering are shown according to the no of links and time per ms per cluster. Both of the K-Mean and K-Medoid are use the same data sets. It is given below:

Dataset1: www.google.com
Dataset2:www.yahoo.com
Dataset3:www.rediffmail.com

Table -1. K-Mean

Data Set	No.of Mails	No. Of clusters	Time per cluster
D1	431	15	5.09 ms
D2	50	20	5.0 ms
D3	635	25	5.8 ms

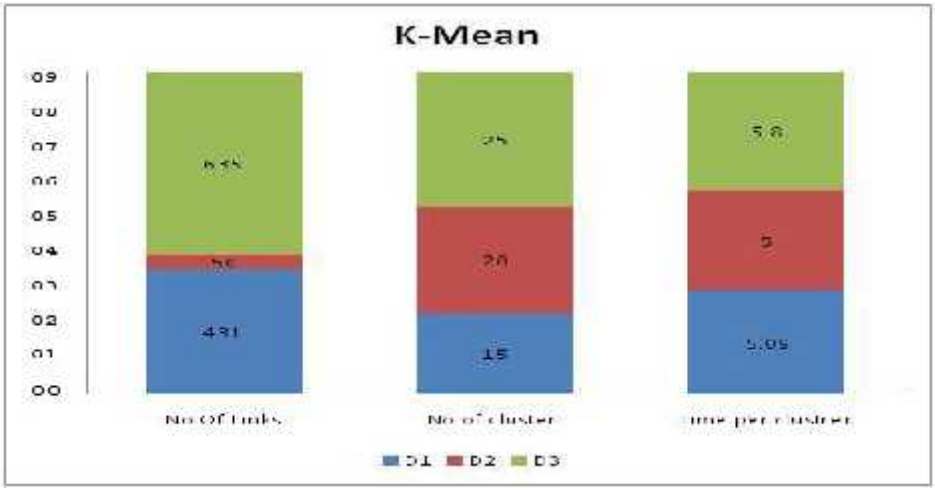

Figure 1. K-Mean

Table 2. K-Medoid

Data Set	No.of Mails	No. of clusters	Time per cluster
D1	431	15	4.6 ms
D2	50	20	4.8 ms
D3	635	25	4.02 ms

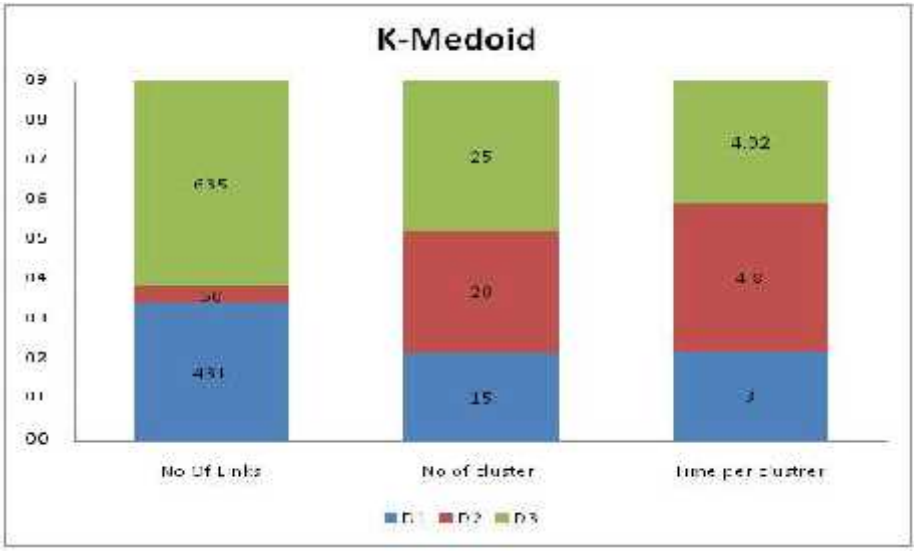

Figure 2. K-Medoid

The graphical representations of both K-mean and K-medoid are shown in which x-axis represents the no. of links and Y-axis representing the time per cluster in ms formed according to no .of links. These two algorithms have implemented source code in the weka 3.7 version based on the average execution time and similarity of the objects. Thees results are summarized in the above tables.

H**ardware Specification**: The above results are taken on Pentium IV processor having 4GB memory and 500GB Hard Disk Drive.

IJMDRR
E- ISSN –2395-1885
ISSN -2395-1877

IV. COMPARISON

This table shows the comparison between K-Means and K-Medoid based on different parameters[10]

Parameters	K-Means	K-Medoid
Complexity	O(i k n)	O(i k(n-k)2)
Efficiency	Comparaticely more	Comparatively less
Implementation	Easy	Complicated
Sensitive to Outliers?	Yes	No
Advance specification of No. Of clusters 'k'	Required	Required
Does intial partition affetcs result and runtime?	Yes	Yes
Optimized for	Separated clusters	Separated clusters, small dataset

Table 3: Comparison between K-Mean & K-Medoid

V. CONCLUSION & FUTURE WORK

In this paper we describe about comaprison between K-Means and K-medoid algorithm. The resulst of the K-Mean and K-Medoid clustering algorithms are shown here.The comparative study of these two algorithms are demonstrating based on the no of custers and also execution time of the links. So we can conclude K-Mediod show the better result with k-Mean. The computational time taken by K-Means is more comapre with K-Medoid accoring to the clusters. As a future work, some other adavnced algorithm can be applied for a better result.

VII. REFERENCES

1. K-Medoid Clustering Algorithm- A Review. International Journal of Computer Application and Technology (IJCAT) Volume 1 Issue 1 (April 2014) ISSN: 2349-1841.
2. R. Suguna,D. Sharmil,, "Clustering Web Log Files – A Review",International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 4, April – 2013.
3. Mrs. G. Sudhamathy, Dr. C. Jothi Venkateswaran, **"**Web Log Clustering Approaches – A Survey", International Journal on Computer Science and Engineering (IJCSE), Vol. 3 No. 7 July 2011.
4. Dr.T Velmurugan , "Efficiency of k-Means and K-Medoids Algorithms for Clustering Arbitrary Data Points",Int.J.Computer Technology & Applications,Vol 3 (5), 1758-1764.
5. Ji Wentian, Guo Qingju, Zhong Sheng , **"**Improved K-medoids Clustering Algorithm under Semantic Web" International Conference on Computer Science and Electronics Engineering (ICCSEE 2013).
6. Hae-Sang Park, Jong-Seok Lee and Chi-Hyuck Jun ," A K-means-like Algorithm for K-medoids Clustering and Its Performance".
7. Jinhua LI, Hengxiang TIAN, Dandan XING ,"Clustering User Session Data for Web Applications Test" Journal of Computational Information Systems (2011).
8. T. Velmurugan and T. Santhanam, "Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points" , Journal of Computer Science 6 (3): 363-368, 2010.
9. Radhika Kyadagiri ,Prof. D. Jamuna ,Masthan Mohammed, "An Efficient Density based Improved K- Medoids Clustering algorithm" ,International Journal of Computers and Distributed Systems Vol. No.2, Issue 1, December 2012.
10. Review Paper: A Comparative Study on Partitioning Techniques of Clustering Algorithms, International Journal of Computer Applications (0975 – 8887) Volume 87 – No.9, February 2014.
11. "Data Mining Concept and Techniques" ,2nd Edition, Jiawei Han, By Han Kamber..