International Journal of Computer Application and Technology
(IJCAT)
Volume 1 Issue 1 (April 2014)
ISSN: 2349-1841

# K-Medoid Clustering Algorithm- A Review

Noor Kamal Kaur[1], Usvir Kaur[2], Dr.Dheerendra Singh[3]
[1]Student, Sri Guru Granth Sahib World University, Fatehgarh Sahib, India
kamalpurewal18@gmail.com
[2]Assistant Professor, Sri Guru Granth Sahib World University, Fatehgarh Sahib, India
usvirkaur@gmail.com
[3]Professor, Shaheed Udham Singh College of Engineering.&Technology, India
professordsingh@gmail.com

***ABSTRACT :*** **Web mining is the application of data mining techniques to extract knowledge from Web data,. Web mining is the application of data mining techniques to extract knowledge from web data, including web documents, hyperlinks between documents, us-age logs of web sites, etc. Web servers record and accumulate data about user interactions whenever requests for resources are received. Analyzing the web access logs of different web sites can help understand the user behavior and the web structure, thereby improving the website design. Log record has useful information such as URL, IP address, time and so on. 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 most important research areas in the field of data mining .Clustering is the process of classifying objects into different groups by partitioning sets of data into a series of subsets called clusters. There are two types of clustering algorithms that is k-medoid and k-mean. It is a process of grouping data objects into disjoint clusters so that the data in each cluster are similar, yet different to the other clusters.**

***KEYWORDS-*** *Clustering, K-medoid, log files*

## I. INTRODUCTION

Data Mining (DM) is the extraction of information from large amounts of data to view the hidden knowledge and facilitate the use of it to the real time applications. DM consists of algorithms for data analysis. Some of the major Data Mining techniques used for analysis are Clustering, Association, Classification and etc. Clustering is an effective technique for data analysis.. Most existing methods of clustering can be categorized into three: 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 is a technique to search hidden patterns that exists in datasets. It is a process of grouping data objects into disjoint clusters so that the data in each cluster are similar, yet different to the other clusters. One of the clustering method that minimizes the clustering error is the k-means algorithm. It is attractive, because it is simple and fast. It partitions the input dataset into k clusters. Each cluster is represented by changing centroid(also called cluster centre), starting from some initial values named seed-points. k-Means computes the distances between the inputs (data points) and centroids, and assigns values to the nearest centroid. However, the k-means algorithm is a local search procedure and it is well known that it has main two limitations that is, first one is that the number of the clusters is unknown, and the second is initial seed problem. In web log files, it find difficult to get it directly. Some preprocessing techniques and pattern discovery algorithms are needed for the web logs to get meaningful information. Always it is better to split the data into some groups with respect to parameters. So that , a various clustering algorithms was applied to web log files to group them either user based, link based or session based.

Web Usage Mining deals with the usage details and behavior of the website visitors. These days it becomes an interesting research field for satisfying customer expectations. The navigation details are maintained in the web servers, proxy servers and client machines in the form of Common Log File format .The user's website visiting details are recorded in various sources in CLF format. The web logs are large in size and are not in proper format. So, preprocessing is applied to make the web logs suitable for extracting knowledge. .Soft clustering algorithms (fuzzy clustering) are chosen because of its efficiency and accuracy. K-means clustering algorithm is chosen because of its simplicity, thus it is modified to improve its efficiency. [4].

Clustering analysis is a widely used data mining algorithm which is a process of partitioning a set of data objects into a number of object clusters, where each data object shares the high similarity with the other objects within the same cluster but is quite dissimilar to objects in other clusters. The survey of three web logs clustering approaches focusing on Web Personalization and serve as a source of ideas for people working on personalization of information systems. It proposes the easy, simple, best approach, the Fuzzy Clustering for user behavior pattern discovery [8].

Various researcher uses arbitrarily distributed input data points to evaluate the clustering quality and performance of two clustering algorithms that is k- Means and k-Medoids. To evaluate the clustering quality, the distance between two data points are taken for analysis. The computational time is calculated for each algorithm in order to measure the performance of the algorithms. The experimental results show that the k-Means algorithm yields the best results compared with k-Medoids algorithm. The average execution time of the k-Means algorithm is very less than the k-Medoids algorithm [1].

K-medoids clustering algorithm is very efficient in classifying cluster categories. Based on algorithm analysis and
Selection improvement of centre point K, web model of ontology data set object is set 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].
An algorithm for K-medoids clustering which works like K-means algorithm. The algorithm calculates the distance matrix once and uses this distance matrix for finding new medoids at every step. The algorithm is using real and artificial data. The algorithm takes the less time in computation with comparable performance as compared to the Partitioning Around Medoids[6].
K-medoids algorithm is adopted as a reduction mechanism to partition user session data into a set of clusters. Each cluster represents similar scenarios of user interactions with a web application. Samples of each cluster are selected and constructed into test data for web applications test. The solutions to the key issues of non-numeric data type of user sessions and their dissimilarity definition are described. Simply count the different number of attributes in two client requests, each of which consists of one basic request, none or many name-value pairs. The suite is generated by randomly selecting representative user sessions from each partitioned cluster without reconstructing or rearranging the user sessions data [7].
The algorithms K-Means and K-Medoids were examined and analyzed based on their basic approach. The best algorithm was found out based on their performance. The input data points are generated by two ways that is one by using normal distribution and another by applying uniform distribution. The randomly distributed data points were taken as input to these algorithms and clusters are found out for each algorithm. The execution time for the algorithms in each category was compared. The accuracy of the algorithm was determined during different execution of the program on the input data points. The average time taken by K-Means algorithm is greater than the time taken by K-Medoids algorithm for both cases [11].

## II. K-MEDIOD

The k-Means algorithm has main disadvantage that it is sensitive to outliers since an object with an extremely large value may distort the distribution of data. Instead of taking the mean value of the objects in a cluster as a reference point, a medoid can be used, which is the most centrally located object in a cluster. Thus, the partitioning method can still be performed based on the principle of minimizing the sum of the dissimilarities between each object and its corresponding reference point. This concept forms the basis of the k-Medoids method. The basic strategy of k- Medoids clustering algorithms is to find k clusters in n objects by first arbitrarily finding a representative object (the medoids) for each cluster. Each remaining object is clustered with the medoid to which it is the most similar. The k-Medoids method uses representative objects as reference points instead of taking the mean value of the objects in each cluster. The algorithm takes the input parameter k, the number of clusters to be partitioned among a set of n objects[11]. K-medoid is a classical partitioning technique of clustering that clusters the data set of n objects into k number of clusters. This k: the number of clusters required is to be given by user. This algorithm works on the principle of minimizing the sum of dissimilarities between each object and its corresponding reference point. The algorithm randomly chooses the k objects in dataset D as initial representative objects called medoids. A medoid can be defined as the object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal i.e. it is a most centrally located point in the given data set. Then for all medoid, after every assignment of a data object to particular cluster the new medoid is decided. The problem is K-Medoids does not generate the same result with each run, because the resulting clusters depend on the initial random assignments. It is more robust than kmedoids in the presence of noise and outliers; however it's processing is more costly than the k-medoid method. Lastly, the optimal number of clusters k is hard to be predicted, so it is difficult for a user without any prior knowledge to specify the value of k [3].

A typical k-Medoids algorithm for partitioning based on medoid or central objects is as follows:

```
Input: k: The number of clusters
D: A data set containing n objects
Output: A set of k clusters that minimizes the sum of
the dissimilarities of all the objects to their nearest
medoid.
Method: Arbitrarily choose k objects in D as the initial
representative objects;
Repeat assign each remaining object to the cluster with
the nearest medoid;
randomly select a non medoid object $O_{random}$;
compute the total points S of swaping object $O_j$
with $O_{ramdom}$;
if $S < 0$ then swap $O_j$ with $O_{random}$ to form the
new set of  k medoid;
Until no change;
```

## IV.EXPERIMENTAL RESULTS

**The Experimental results are shown:**

In our Experimental results we consider three data sets named D1, D2, and D3.All these data set consist of log files. We have shown the clustering of log files according to no. of links and time per cluster in ms is also shown. For both K-Mean and K-medoid we have use the same data sets named as follows:

**Dataset1:** www.test.lookscare.com

**Dataset2:** www.travel.stepstoindia.com

**Dataset3**: www.monitorware.com

**K-medoid**

| Data set Name | No of links | No of clusters | Time per cluster |
|---|---|---|---|
| DI | 429 | 10 | 4.09ms |
| D2 | 47 | 20 | 4.0ms |
| D3 | 861 | 25 | 4.8ms |

**Figure**

K- Mean

| Data Set Name | No of links | No of clusters | Time per cluster |
|---|---|---|---|
| **D1** | **429** | **10** | **4.6ms** |
| **D2** | **47** | **20** | **4.02ms** |
| **D3** | **861** | **25** | **5.78ms** |

**Figure**

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.

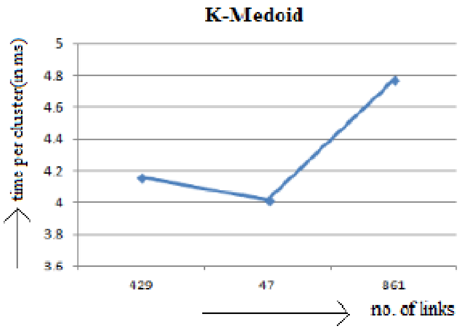

**Figure 1. (a)**

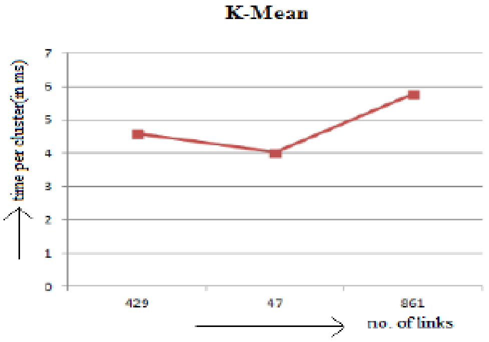

**Figure 2. (a)**

**Hardware Specification**: The above results are taken on Intel Core i3 processor having 4GB memory and 500GB Hard Disk Drive.

**Comparative study**:
Comparison between K-mean and K-medoid

| K-mean | K-medoid |
|---|---|
| Complexity is O(I k n) | Complexity is O(I k(n-k)^2) |
| Sensitive to outliers | Not Sensitive to outliers |
| Implementation of algorithm is easy | Implementation of algorithm complicated |
| Less robust | More robust |
| Execution time per cluster is more | Execution time per cluster is less |

## V.CONCLUSION

In this paper we concluded that k-medoids demonstrate better performance than k-means. We have shown the results of both k- mean and k-medoid clustering algorithms. The comparative study of both clustering algorithm are shown according to the no. of clusters formed according to no. of links and execution time taken by both clustering algorithm to make the clusters. The computational time taken by K-mean to cluster according to number of links is more as compared to k-medoid.

## V1.REFERENCES

[1].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

[2].V.CHITRAA, Dr. ANTONY SELVADOSS THANAMANI, “An Enhanced Clustering Technique for Web Usage Mining,International“ , Journal of Engineering Research & Technology (IJERT)Vol. 1 Issue 4, June - 2012.

[3].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

[4].R. Suguna,D. Sharmil,, "Clustering Web Log Files – A Review",International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 4, April - 2013.

[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]. 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.

[9]. ENG TALAL ALSULAIMAN, “Classifying Technical Indicators Using K-Medoid Clustering”.

[10]. A. P. Reynolds, G. Richards, and V. J. Rayward-Smith ,“The Application of K-medoids and PAM to the Clustering of Rules”.\

[11]. 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.