ISSN:2249-5789
Sachin Shinde et al , International Journal of Computer Science & Communication Networks,Vol 4(6),197-202

# Improved K-means Algorithm for Searching Research Papers

Sachin Shinde
*Department of Computer Engineering*
*Flora institute of Technology, Pune*
*Maharashtra, India*
*Sachinsss2986@gmail.com*

Bharat Tidke
*Department of Computer Engineering*
*Flora institute of Technology, Pune*
*Maharashtra, India*
*Sachinsss2986@gmail.com*

## Abstract

*Clustering is one of the unsupervised learning method in which a set of essentials is separated into uniform groups. The k-means method is one of the most widely used clustering techniques for various applications. For the Searching as well as reading research papers users need more time or users spend two to three hours for searching or reading single papers, so this is more consuming process, so it is required that use enhanced search engine which is based on fastest reading algorithm which provides best output or results. So we are proposed Enhanced architecture with improved k-means algorithm, which proposes a method for making the algorithm more effective and efficient, so as to get better clustering with reduced complexity. It will search the base keyword of the content from the knowledge database. Proposed work uses the search engine based on clustering and text mining.*

**Keywords**-*Text mining, Clustering, K-means Algorithm, Enhanced K-means Algorithm.*

## 1. Introduction

Data mining, a synonym to "knowledge discovery in databases" is a process of Analyzing data from different perspectives and summarizing it into useful information. Clustering [2] is useful technique for the discovery of data distribution and patterns in the underlying data. Clustering is an example of unsupervised classification. Classification refers to a procedure that assigns data objects to a set of classes. Unsupervised Classifications means that clustering does not depend on predefined classes and no external teacher set is used.

.

The use of search engines to locate information has grown rapidly supported on the needs of users generating a snowball significance, where all the assemblage is obtainable in dissimilar websites, including information that is not useful or significant but also included information of scientific interest to remember that not all the information is by default in specialized research journals, transactions, letters and magazine in the area of computers science or advances in technology like Springer, IEEE, ACM or papers recognized of research and more.

The implementation of a better tool to investigate research articles, it would be useful to the conclusion and minimize times of investigation or search and it also make a best engine to get best results in every search of research papers. The use of K-Means algorithm allow us to implement semi-supervised learning clusters using an algorithm so as to help to recognize approximate the text to search using predefined patterns and the implementation of a cluster algorithm for consultations within the database manager MySQL i.e. database manager that allows free use of multithreading, multi-search and multi-user in order to obtain scientific research papers. But for large data sets the computational complexity of the original k-means algorithm is very high. So finally the algorithm results in different types of clusters depending on the random choice of initial centroids. Researchers made several attempts to improve K-means Algorithm. This paper deals with a method for improving the accuracy and efficiency of the k-means algorithm. Cluster analysis is a one of the primary data analysis tool in the data mining. Clustering algorithms are mainly categorized into two types: Hierarchical algorithms and Partition algorithms. A hierarchical clustering algorithm divides the given data set into

smaller subsets in hierarchical fashion. A partition clustering algorithm separates the data set into desired number of sets in a single step [9]. Numerous methods have been proposed to solve clustering problem.

This paper is organized as follows. Section 2 presents an overview of k-means algorithm and a short analysis of related Work. Section 3 introduces proposed method including Architecture and Algorithm. Section 4 describes the conclusion and References.

## 2. Traditional k-means Algorithm

### A. Introduction

K-means clustering [8] is a partition-based cluster analysis method. According to this algorithm we firstly select k data value as initial cluster centers, then calculate the distance between each data value and each cluster center and assign it to the closest cluster, update the averages of all clusters, repeat this process until the criterion is not match. K-means clustering aims to partition data into k clusters in which each data value belongs to the cluster with the nearest mean. Figure 1 shows how to process of the basic K- means.

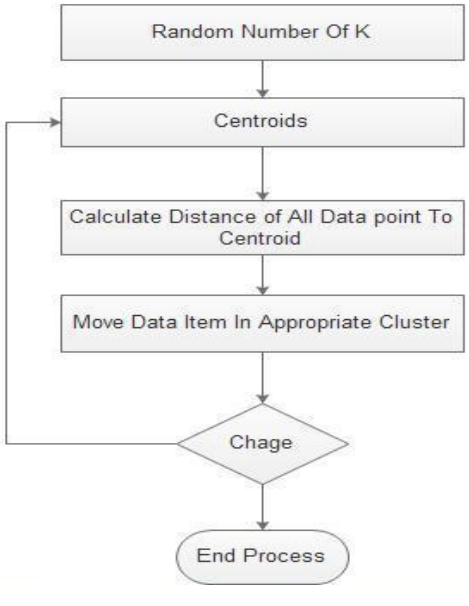


Fig: K-means Algorithm Process

**Steps:**

Step 1.Begins with a decision on the value of k= number of clusters.

Step 2. Put any initial partition that classifies the data into k clusters. You may assign the training samples randomly, or systematically as the following: Take the first k training sample as single-element clusters assign each of the remaining (N-k) training sample to the cluster with the nearest centroid.After each assignment, recomputed the centroid of the gaining cluster.

Step 3 .Take each sample in sequence and compute its distance from the centroid of each of the clusters. If a sample is not currently in the cluster with the closest centroid, switch this sample to that cluster and update the centroid of the cluster gaining the new sample and the cluster losing the sample.

Step 4 . Repeat step 3 until convergence is achieved, that is until a pass through the training sample causes no new assignments. If the number of data is less than the number of cluster then we assign each data as the centroid of the cluster. Each centroid will have a cluster number. If the number of data is bigger than the number of cluster, for each data, we calculate the distance to all centroid and get the minimum distance. This data is said belong to the cluster that has minimum distance from this data. Since we are not sure about the location of the centroid, we need to adjust the centroid location based on the current updated data. Then we assign all the data to this new centroid. This process is repeated until no data is moving to another cluster anymore.

The algorithm works by using the following equation

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2 \quad \text{.....(1)}$$

The formula represents a given set of observations (X1,X2… Xn ) where each observation represent an element of the cluster with a d-dimensional real vector , k-means clustering aims to partition and the n observations into k sets (k <= n ) S = { S1, S2, ….. Sk ) that's to minimize the cluster where μi is the mean of points in Si.

### B. Related Work

There is several techniques were implemented by researchers to improve the effectiveness and efficiency of k-means algorithm. Several methods have been proposed in the literature for finding the better initial centroids. And some methods were proposed to improve both the accuracy and efficiency of the k-means clustering algorithm.

M. Fahim et al. [1] proposed an enhanced method for assigning data points to the suitable clusters. In the original k-means algorithm in each iteration the distance is calculated between each data element to all centroids and the required computational time of this algorithm is depends on the number of data elements, clusters and iterations, so it is computationally cheap. In Fahim approach the data elements are assign to the appropriate clusters to reduce computational time. But in this appoch the initial centroids are selected randomly or deliberately. So this approach is very sensitive to the initial starting points and it does not guarantees to produce the unique clustering results.

K. A. Abdul Nazeer et al. [2] proposed an enhanced algorithm to improve the accuracy and efficiency of the k-means clustering algorithm. In this algorithm two methods are used, i.e finding the better initial centroids and efficiently assigning data points to appropriate clusters with reduced time complexity. So in this algorithm we are produces good effective clusters in minimum computational time.

Zhang Chen et al. [3] proposed the initial centroids algorithm based on k-means that have avoided alternative randomness of initial center.

Fang Yuan [4] proposed the initial centroids algorithm where initial centroids are calculated systematic way. If different initial values are given for the centroids, the accuracy output by the standard k-means algorithm can be affected on selecting cluster approach.

Koheri Arai et al. [5] proposed an algorithm where both k-means and hierarchical algorithms are used to find out better initial cluster center for k-means.

Bhattacharya et al. [6] proposed DCCA which is a novel clustering algorithm, called Divisive Correlation Clustering algorithm which is used to produce clusters without taking the initial centroids and k value which is assign as number of cluster as a input.

## 3. Proposed Method

### A. Architecture

**Search Architecture Model:**

The implementation of data mining to solve a problem involves the need to implement a methodology focus into the analysis of pattern into the texts, where there are several methodologies tailored built-oriented identify of attributes that will be reviewed, so this kind of methodologies are not used for our implementation as the proposed work need a methodology to be acceptable evolutionary behaviour.

In figure 1 shows the search architecture model [9], in search architecture model the text mining is mainly used and as a first part of the architecture as a problem outline where the user starts a searching through various interfaces for entering text within the information used in the process of searching patterns within a knowledge base for accessing parameters for the selection of cluster where the actual search was implemented, once the searching is completed within the database in the process of localization papers.

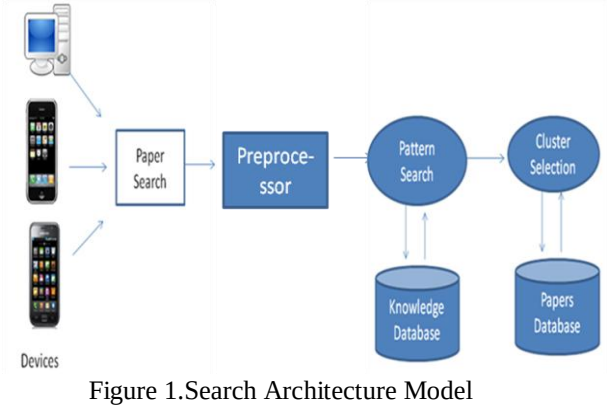


Figure 1.Search Architecture Model

In search architecture we will inputs the location of the research to get text patterns and work reading line by line and in this process is supported by a knowledge portion that is fed into a rank semi-automatically with the information collected from items previously stored. Before searching the pattern of knowledge we will apply pre-processing for reducing time and size complexity of process, so we use stop word elimination approach. The uses the search pattern into the research paper for accessing the database searching patterns of knowledge, with this the proposed work generate there engine to make pattern matching needed in the problem

outline start with a pattern matching that read the article searching a similar pattern compared with the knowledge base once we are locate in where it relates is selected cluster on which will be uploaded from the article in the database.

**Pattern matching Architecture model:**
In the pattern Matching architecture [9] figure 2 for looking to the database searching patterns of knowledge it uses the search pattern into research papers, with this the proposed work we are generating engine to make pattern matching needed in the problem outline start with a pattern matching that read the article searching a similar pattern compared with the knowledge base once we will locate in which it relates is selected cluster on which will be uploaded from the article in the database.

The use of clusters involves the categorization of several groups partitioned that get a distinctive in this way determines a measure of characteristics between the stored collection in the knowledge database.

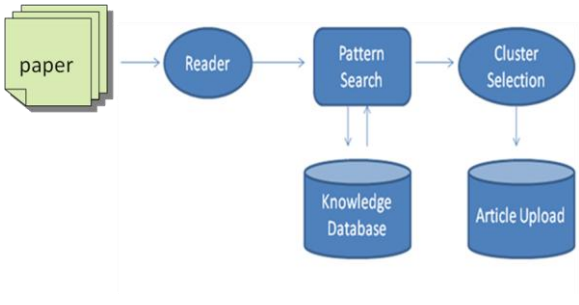


Figure 2.Pattern matching architecture model

This type of behaviour or conduct can be categorized as clustering semi-automatic learning, which depends on the classification algorithm, is implemented and the type of measures used to feed because it depends on the type of information or parameters are assigned and that adding a number undefined variables can degrade the performance of the implemented algorithm because all attributes are not relevant to the classification of information. The process should select those that represent relevant proper values for effective classification.

## B. Improved K-means Algorithm

In improved k-means algorithm we are enhancing the performance of k-means clustering algorithm. In this method the initial centroids are selected randomly, So this method is very sensitive to the initial starting points and it does not guarantees to produce the unique clustering results. In the paper [2]authors uses two methods for finding initial clustering i.e finding initial centroids and assingning data points to appropriate clusters. so we are proposed an enhanced algorithm to improve the accuracy and efficiency of the k-means clustering algorithm. In this paper we are proposed a new approach for finding the better initial centroids with minimized time complexity. In the proposed algorithm first we will checking, the given data set contain the negative value attributes or not. If the data set contains the negative value attributes then we are transforming the all data points in the data set to the positive attribute value in the given data set. Here positive space is subtracting the each data point attribute with the minimum attribute value in given data set. Transformation is required, because in the proposed algorithm we will calculate the distance from origin to each data point in the data set. So, when we are selecting the different data points then we will get the same Euclidean distance from the origin. Because of that we will get result is in incorrect selection of the initial centroids. So to overcome this problem all the data points are transformed to the positive space. Then for all the data points as we will get the unique distances from origin. If there is all positive values in data set then the transformation is not required.
In the next step of algorithm, for each data point we will calculate the distance from origin. Then, the original data points are sorted accordance with the sorted distances. After sorting partition the sorted data points into k equal sets or numbers. In each set take the middle points or mean value as the initial centroids. These initial centroids lead to the better unique clustering results. Next, for each data point the distance calculated from all the initial centroids. The next stage is an iterative process which makes use of a heuristic approach to reduce the required computational time. The data points are assigned to the clusters having the closest centroids in the next step. ClusterId of a data point denotes the cluster to which it belongs. NearestDist of a data point denotes the present nearest distance from closest centroid.

**Algorithm : The Enhanced Method**

Require: D = {d1, d2, d3,..., di,..., dn } // Set of n data points.

di = { x1, x2, x3,..., xi,..., xm } // Set of attributes of one data point.
k // Number of desired clusters.
Ensure: A set of k clusters.
Steps:
1: In the given data set D, if the data points contains the both positive and negative attribute values then go to step otherwise go to step 4.
2: Find the minimum attribute value in the given data set D.
3: For each data point attribute, subtract with the minimum attribute value.
4: For each data point calculate the distance from origin.
5: Sort the distances obtained in step 4. Sort the data points accordance with the distances.
6: Partition the sorted data points into k equal sets.
7: In each set, take the middle point as the initial centroid.
8: Compute the distance between each data point di (1 <= I <= n) to all the initial centroids cj (1 <= j <= k).
9: Repeat
10: For each data point di, find the closest centroid cj and assign di to cluster j.
11: Set ClusterId[i]=j. // j:Id of the closest cluster.
12: Set NearestDist[i]= d(di, cj).
13: For each cluster j (1 <= j <= k), recalculate the centroids.
14: For each data point di,
14.1 Compute its distance from the centroid of the present nearest cluster.
14.2 If this distance is less than or equal to the present nearest distance, the data point stays in the same cluster.
Else
14.2.1 For every centroid cj (1<=j<=k) compute the distance d(di, cj).
End for;

Until the Relevant criteria is met.

In the next step, for each cluster the new centroids are calculated by taking the mean of its data points. Then for each data point the distance calculated from the new centroid of its present nearest cluster. If this distance is less than or equal to the previous nearest distance, then the data point stays in the same cluster, otherwise for each data point we are need to calculate the distance from all centroids. After calculated the distances, the data points are assigned to the appropriate clusters and thenew ClusterId's are given and new NearestDist values are updated. This reassigning process is repeated until the convergence criterion is met.

## Conclusion

The proposed algorithm is to be more accurate and efficient compared to the original k-means algorithm. This proposed method finding the better initial centroids and provides an efficient way of assigning the data points to the suitable clusters. So it is very effective to accessing research papers using Enhanced clustering algorithm.

## References

[1] A. M. Fahim, A. M. Salem, F. A. Torkey and M. A. Ramadan, "An Efficient enhanced k-means clustering algorithm," journal of Zhejiang University, 10(7): 16261633, 2006.

[2] K.A.Abdul Nazeer and M. P. Sebastian, "Improving the accuracy and efficiency of the k-means clustering algorithm," in International Conference on Data Mining and Knowledge Engineering (ICDMKE), Proceedings of the World Congress on Engineering (WCE-2009), Vol 1, July 2009, London, UK.

[3] Chen Zhang and Shixiong Xia, " K-means Clustering Algorithm with Improved Initial center," in Second International Workshop on Knowledge Discovery and Data Mining (WKDD), pp. 790-792, 2009.

[4] F. Yuan, Z. H. Meng, H. X. Zhangz, C. R. Dong, " A New Algorithm to Get the Initial Centroids," proceedings of the 3rd International Conference on Machine Learning and Cybernetics, pp. 26-29, August 2004.

[5] Koheri Arai and Ali Ridho Barakbah, "Hierarchical K-means: an algorithm for Centroids initialization for k-means," department of information science and Electrical Engineering Politechnique in Surabaya, Faculty of Science and Engineering, Saga University, Vol. 36, No.1, 2007..

[6] A. Bhattacharya and R. K. De, "Divisive Correlation Clustering Algorithm (DCCA) for grouping of genes: detecting varying patterns in expression profiles," bioinformatics, Vol. 24, pp. 1359-1366, 2008.

[7] Elmasri, Navathe, Somayajulu, Gupta, Fundamentals of Database Systems, Pearson Education, First edition, 2006.

[8] S. Deelers and S. Auwatanamongkol, "Enhancing K- Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance," International Journal of Computer Science, Vol. 2, Number 4

[9] E.AlanCalvillo,AlejandroPadilla,JaimeMunoz"SearchinsResearchpapers using clustering and text minig",IEEE 2013.

[10] Masip, P. Busqueda de Informacion academica en Internet.2009;Availablefrom:http://www.slideshare.net/p.masip/buscadores-acadmicos-3052335 last consult 20/02/2012.

[11] J. A. Olivas, J.d.l.M., J. Serrano-Guerrero, P. J. Garces, F. P. Romero, Desarrollo de Motores Inteligentes de busqueda en Internet en el Marco del grupo de Investigacion SMILe-ORETO. 2006. ISBN 84-9750-525-5.

[12] Chakrabarti, S., Mining The Web: Discovering knowledge from hypertext data, Part 2. 2003.