Systematic Selection of Initial Centroid for K-Means Document Clustering System

Tin Thu Zar Win[#1], Moe Moe Aye[*2]

[#1] Department of Computer Engineering and Information Technology, Mandalay Technological University
Mandalay, Republic of the Union of Myanmar
[1] zarzar84mtu@gmail.com
[2] moeaye255@gmail.com

Abstract – **As the number of electronic documents generated from worldwide source increases, it is hard to manually organize, analyze and present these documents efficiently. Document clustering is one of the traditionally data mining techniques and an unsupervised learning paradigm. Fast and high quality document clustering algorithms play an important role in helping users to effectively navigate, summarize and organize the information. K-Means algorithm is the most commonly used partitioned clustering algorithm because it can be easily implemented and is the most efficient one in terms of execution times. However, the major problem with this algorithm is that it is sensitive to the selection of initial centroid and may converge to local optima. The algorithm takes the initial cluster centre arbitrarily so it does not always guarantee good clustering results. Different initial cluster centres often lead to different clustering and thus provide unstable clustering results. To overcome this problem, Systematic Selection of Initial Centroid for K-Means (SSIC K-Means) approach is proposed to improve the quality of clustering in this paper. Unlike the traditional K-Means clustering, the proposed SSIC K-Means method can generate the most compact and stable clustering results based on maximum distance initial centroids points instead of random initial centroid points. In this paper, experimental results are presented in F-measures using 20 Newsgroup standard datasets. The evaluations demonstrate that the proposed solution outperforms the other initialization methods and can be applied for other various standard datasets.**

Keywords – **Document clustering, Data mining, K-Means, Initial centroid, SSIC K-Means**

I. INTRODUCTION

In today's world, as the development of new smart technologies, the world is going digital data over the large network. Digital collections of data continue to grow exponentially as the information age continues to infiltrate every aspect of society. To organize these numerous data, clustering system is a novel solution which can be utilized to automatically group related content together. Initially, Document clustering techniques have been receiving more and more attentions as a fundamental and enabling tool for efficient organization, navigation, retrieval, and summarization of huge volumes of text documents. Each cluster contains objects that are very similar to each other and very dissimilar to objects in other clusters. Document clustering has been studied intensively because of its wide applicability in areas such as web mining, search engines, information retrieval, and topological analysis [1].

With a good document clustering method, computers can automatically organize a document corpus into a meaningful cluster hierarchy, which enables an efficient browsing and navigation of the corpus. An efficient document browsing and navigation is a valuable complement to the deficiencies of traditional IR technologies [2].

Clustering can be recognized as the unsupervised classification of patterns (observations, data items, or features vectors) into groups (clusters). Data Clustering can be broadly categorized into partitioning method, hierarchical method, density based method, fuzzy clustering method, artificial neural clustering method, statistical clustering method, grid based clustering method and so on. In these approaches, partitional and hierarchical clustering algorithms are two key approaches in research communities. Partitional clustering aspires to directly acquire a single partition of the set of items into clusters. Most of these approaches are based on the iterative optimization of a criterion function depicting the "agreement" between the data and the partition. Partitional clustering algorithms partition the data set into a particular number of clusters and then evaluate them on the basis of a criterion [3]. There have been several variants of partitioning clustering algorithm.

The most extensively employed partitional algorithm is the iterative K-Means approach. The K-mode algorithm extends the K-Means algorithm by using simple matching dissimilarity for categorical objects, modes instead of means for clusters and frequency based method to update modes in the clustering process. Then, K-Medoid or Partitioning Around Medoid (PAM) deals with the problem of outlier posed by K-Means. K-medoid is similar to K-mean except that mean of each cluster is the value which is nearest to the center of cluster. CLARA is an extension of K-medoid dealing with results of large datasets as K medoid cannot deal with the same. Fuzzy K-Means is another partitional algorithm that uses the concept of degree of membership in each cluster. Among these variants, K-Means is popular partitional clustering algorithm due to the ease of implementation, simplicity, efficiency, and empirical success [4]. The major drawback of this algorithm is that it produces different clusters for different sets of values of the initial centroids. Quality of the final clusters heavily depends on the selection of the initial centroids. There have been many modified versions of K-Means algorithm on clustering with various standard datasets. Improving K-Means algorithm is computationally expensive and requires time proportional to the product of the number of data items, number of clusters and the number of iterations. So, many researchers try to be an optimal K-Means clustering method by modifying several efficient ways [5].

In this paper, the enhanced K-Means algorithm, SSIC K-Means, is presented. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because different location causes different result. So, the better choice is to place them as much as possible far away from each other. Instead of randomness initial centroid of original K-Means, the

proposed algorithm starts by finding the maximum distance objects as initial centroids. Therefore, SSIC K-Means algorithm can generate more accurate cluster results with less computational complexity than simple K-Means algorithm.

The rest of the paper is organized as follows: the literature of this work is reviewed in Section II. The details of background theory are discussed in Section III. The proposed model of this system is briefly explained in Section IV. The performance evaluation measures are presented in Section V. The paper is concluded in section VI.

II. Related Work

In data mining, clustering plays an important role for finding data information and pattern recognition. Among different clustering algorithms, K-Means algorithm is undoubtedly the most widely used partitional clustering algorithm. However, K-Means has two significant disadvantages. First, it is sensitive to the outlier and noise. Second, it is highly sensitive to the selection of the initial clusters. Adverse effects of improper initialization include empty clusters, slower convergence, and a higher chance of getting stuck in bad local minima. To address the initialization problem, different methods have been proposed especially for high dimensional data [6].

Recently, C. Xuhui, and X. U. Yong [7] presented a modified K-Means clustering approach to improve the efficiency of clustering algorithm. Their system used the traditional data space density approach which chooses high-density of objects within the regions as the initial cluster center. The fundamental task is to accurately define high density region. So, it eliminates the randomness of initial cluster center of K-Means algorithm. However, the system leads to clustering result instability if the density parameters, neighborhood radius and experience-values are not correctly defined for density calculation.

D. J. Kalpana and P. S. Nalwade [8] proposed a modified K-means algorithm which has additional steps for selecting better cluster centers. The system computes Min and Max distance for every cluster and finds high density objects for selection of better k. It gives effective results for K-means algorithm by splitting dataset and selects high dense object as cluster centers. But, this system increases the computational complexity and recalculates cluster centers in much iteration.

O. A. Mohamed Jafar and R. Sivakumar [9] presented a hybrid K-Mean++ with Particle Swam Optimization technique (K++-PSO) clustering algorithm. In this system, PSO is used which is one of the evolutionary algorithms for solving clustering problems. It proposes K++-PSO algorithm by using different distance metrics including City Block and Chebyshev distance metrics. The better clustering result is produced in terms of their fitness function value as compared to other algorithms: K-Means, K-Mean++, and K-Medoids. The result shows that it has good performance result for Chebyshev distance than other distance metrics.

S. Mubeena, et al. [10] presented an optimizing data clustering method by using modified K-means clustering algorithm. Initial centroids of clusters are determined systematically by finding closest data-point sets. Firstly, initial centroids are obtained by averaging all the vectors in each data-point set. After that, these data-points are assigned to the clusters which have the closest centroids and then the centroids are recalculated until the convergence criteria are met. It can be applied to more complex data matrices with high dimensionality.

G. Navjot and C. Tejalal [11] proposed a high dimensional clustering scheme for data classification. In this research, the K-Means algorithm is modified for finding the better and stable performance of clustering. The improvement is made on the traditional K-Means clustering approach by implementing Genetic algorithm. The Genetic algorithm uses three main concepts for the selection of initial centroids: reproduction, natural selection and diversity of the genes. The genetic algorithm works till entire data is evaluated. Evaluated results show the performance of their clustering algorithm is optimum and less fluctuating as compared to the traditional clustering algorithm. But, the only limitation in their technique is that it consumes more time and memory as compared to the traditional clustering algorithm.

In this paper, an efficient document clustering system using SSIC K-Means is presented. Instead of randomness initial centroid of simple K-Means algorithm, SSIC K-Means finds furthest points of data objects based on Euclidean distance metric of all documents. Firstly, the proposed method can generate the first two initial points for k cluster and then calculates other initial points by averaging the first and next furthest points. It recalculates the average maximum points until the number of k cluster. Therefore, this method can solve the initial cluster problem of simple K-Means and produce more optimal cluster results for various standard datasets with very high dimensions.

III. Background Theory

A. Clustering

Clustering is the process of finding groups of objects such that the objects in a group will be similar to one another and different from the objects in other groups. Cluster analysis is one of the major data analysis methods widely used for many practical applications in emerging areas. A good clustering method will produce high quality clusters with high intra-cluster similarity and low inter-cluster similarity. The quality of a clustering result depends on both the similarity measure used by the method and its implementation and also by its ability to discover some or all of the hidden patterns [12].

B. K-Means Clustering

K-Means clustering is one of the unsupervised computational methods used to group similar objects in to smaller partitions called clusters so that similar objects are grouped together. The algorithm aims to minimize the within cluster variance and maximize the intra cluster's variance. The K-means clustering algorithm is one of the simplest clustering algorithms in which the number of clusters to be grouped is fixed a priori by the user. The algorithm proceeds by randomly defining *k* centroids and assigning a document to the cluster that has the nearest centroid to the document. Then, for every data point, the minimum distance is determined and that point is assigned

to the closest cluster. This step is called cluster assignment, and is repeated until all of the data points have been assigned to one of the clusters. Finally, the mean for each cluster is calculated based on the accumulated values of points in each cluster and the number of points in that cluster. Those means are then assigned as new cluster centroids, and the process of finding distances between each point and the new centroids is repeated, where points are re-assigned to the new closest clusters. The process iterates for a fixed number of times, or until points in each cluster stop moving across to different clusters. This is called convergence [13].

Euclidean metric is considered as it is one of the widely used distance metrics incorporated with K-means clustering and one that is easy to implement. Also it results in a best solution. Euclidean distance is given in equation 1:

$$\text{dist}(P, C) = \sqrt{\sum_{i=1}^{n} (P_i - C_i)^2} \quad (1)$$

Where P is the data point, C is the cluster center, and n is the number of features. The flowchart of the K-Means algorithm is as shown in Fig. 1.

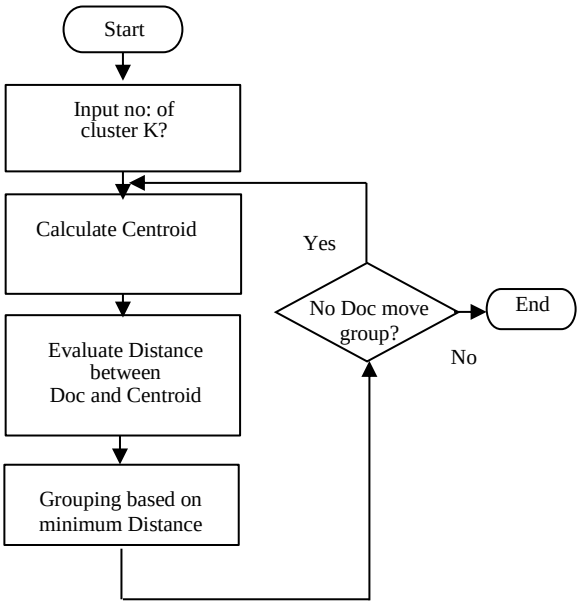

Fig. 1 Flow Chart of K-Means Algorithm [14]

The K-Means algorithm is composed of the following steps:

Algorithm 1: K-Means Clustering Algorithm [14]

Input: Number of desired clusters, k, and a database D={d1,d2,…dn} containing n data objects.

Output: A set of k clusters

Steps:

(1) Here we have to select randomly k data objects from dataset D as initial cluster centers.

(2) Repeat;

(3) Then calculate the distance between each data object di (1 <= i<=n) and all k cluster centers cj(1<=j<=k) and assign data object di to the nearest (closest) cluster.

(4) For each cluster j, recalculate the cluster center.

(5) Until no changing in the center of clusters.

IV. The Proposed Clustering Model

The idea of the proposed model is to cluster the documents by modifying the initial centroid selection method of K-Means algorithm. The proposed method, SSIC K-Means, eliminates the random initial centroid selection of simple K-Means algorithm. Unlike simple K-Means algorithm, it is based on determining maximum dissimilarity occurrence of all documents. In the proposed method, the procedure is almost similar to the original K-Means algorithm except that the initial centroids are computed systematically. The proposed system mainly consists of five main phases: document collection, document pre-processing, initial centroid selection, calculating distance between centroid and documents, and document clustering based on minimum distance.

The flow chart of the proposed model is shown in Fig.2.

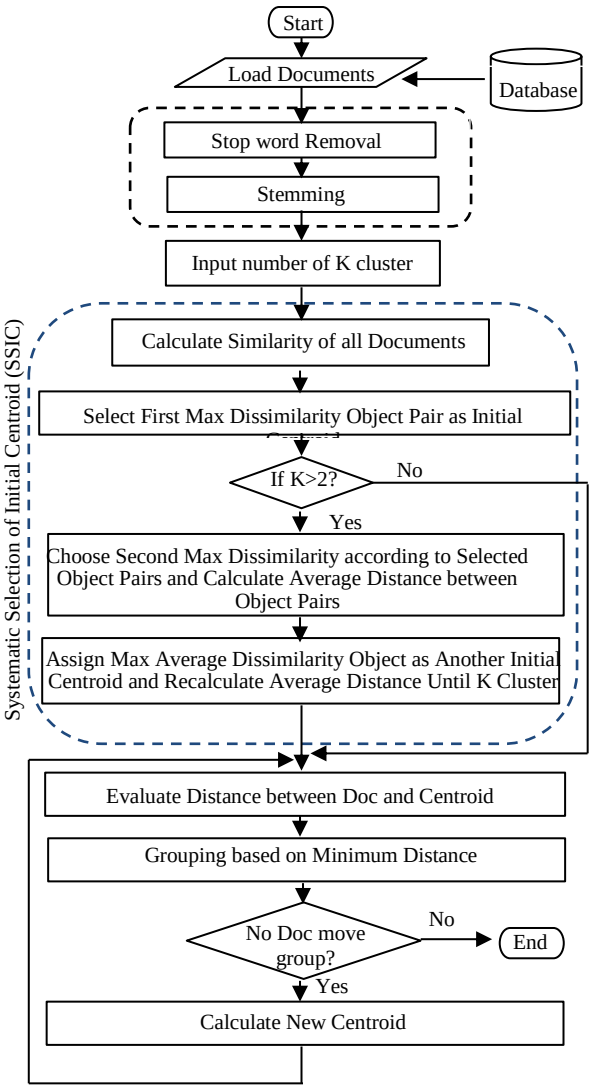

Fig. 2 The procedures of the proposed model

In document collection, 1000 documents were collected from 20 NewsGroup datasets. These documents undergo refinement which is fed to the algorithm to obtain clusters containing documents from similar domains.

In document pre-processing, two techniques namely, removing stop words and stemming algorithm are used. The pre-processing is a process to optimize the list of terms that act as the keywords list. The concept of pre- processing is used to prune all character and term from the document with poor information. Pre-processing is a very important

step since it can affect the result of a clustering algorithm. So, it is necessary to preprocess the data sensibly.

Removing stop words starts the first process. The stop words are words that carry no information and meaning less when we use them as search term (keyword) (i.e., Pronouns, prepositions, Conjunctions, Punctuations). Stop words may be eliminated using a list of stop words. Stop words removal removes stop (linking) words like “have”, “then”, “it”, “can”, “need", “but”, “they”, “from”, “was”, “the”, “to” and “also” from the documents.

The second process is stemming a word. Stemming is the process of reducing words to their stem or root form. Stemming also removes the prefixes and suffixes of each word. For example ‘cook’, ‘cooking’, ‘cooked’ are all forms of the same word used in different constraint but for measuring similarity these should be considered same [15].

In initialization step, the proposed SSIC K-Means includes three main processes. At the first step, the proposed method constructs the similarity matrix of all documents which undergo all refinement steps. Euclidean distance metric is used to calculate the similarity distance of all documents. From these results, the first two furthest data points are assigned as initial centroids. And then, other initial points are determined based on average distance between furthest data points and assign maximum average distance object as the other initial centroid. So, the proposed method recalculates average maximum object until the number of k-cluster.

In the next step, the procedure is almost similar to the original K-Means algorithm except that the initial centroids are computed systematically. It calculates the distance between each document and initial centroid and assign data object with minimum distance to the nearest (closest) cluster. For each cluster, the centroids are recalculated by taking the mean of the values of its data-points. So, this stage is an iterative process which recalculates the cluster centers until no more change the cluster results. This paper proposes a systematic approach to determine the initial centroids so as to produce clusters with better accuracy. The proposed SSIC K-Means method is outlined as Algorithm 2.

Algorithm 2: Proposed SSIC K-Means Algorithm

Input: Number of desired clusters, k, and a database D = {d1,d2,…dn} containing n data objects.
Output: A set of k clusters
Steps:
(1) Initially, calculate similarity distance metric of K data objects from dataset D.
(2) Select the first maximum dissimilarity object pair from dataset D as initial cluster centres
(3) If (k>2) then
  (i) Choose the second maximum dissimilarity according to selected object pairs and calculate the average distance between object pairs.
  (ii)Assign maximum average dissimilarity object as another initial centroid.
  Else go to step (5).
(4) Repeat step (3) until to K cluster.
(5) Calculate the distance between each data object di (1<= i<=n) and all k cluster centers c_j (1<=j<=k) and assign data object d_i to the nearest (closest) cluster.
(6) For each cluster, recalculate the cluster center.
(7) Until no changing in the center of cluster.

V. Experimental Results

In this section, the experimental results are presented. The experiment is carried on 20_NewsGroups dataset by gradually increasing number of documents from 100 to 1000. The 20 _ NewsGroup dataset is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. The 100 documents are randomly selected from ten categories and developed the data set 20ns which consists of 1000 documents. The F-measure is a combination of precision and recall values used in information retrieval and document clustering systems. So, F-measure is used to test the effectiveness of K-Means and SSIC K-Means clustering approaches by applying on ten different class label datasets of 20-NewsGroup. The values of F-measure are shown in table 1 and the larger value of F-measure is better.

TABLE I
The Values of F-Measure

Class	Cluster	Documents	F-measure	
			K-Means	SSIC K-Means
3	3	300	0.527	0.621
5	5	500	0.493	0.550
7	7	700	0.500	0.518
10	10	1000	0.396	0.462

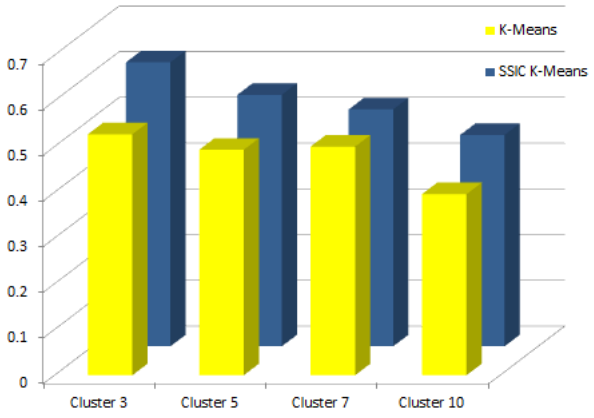

Fig. 3 Comparison of F-Measure between K-Means and SSIC K-Means on 20_ NewsGroup datasets

The comparison of F-measure calculated with SSIC K-Means approach and simple K-Mean approach is shown in Fig. 3. It can be clearly seen that value of F- Measure for traditional K-Means method is less than the SSIC K-Means method for each dataset. The value of F-measure using SSIC K-Means approach is 0.621, so it is near to 1 and simple K-Means approach is 0.527. It may be noted that the proposed method outperforms the traditional K-Means in terms of cluster quality.

VI. Conclusions

K-Means algorithm is widely used for clustering large datasets. However, it does not always guarantee good results as the accuracy of the final clusters depends on the selection of initial centroids. In this paper, an efficient K-Means algorithm for document clustering based on a novel

centroid selection method is presented. The proposed SSIC K-Means algorithm eliminates the random centroid selection approach by replacing systematic centroid selection approach. So, the proposed method produces significantly better clustering solutions. The proposed approach is helpful in selecting significant centers for K-Means clustering. From the experiments on the datasets, it is observed that the proposed method outperforms the traditional K-Means algorithm in term of cluster quality and can be applied for unsupervised clustering of various datasets.

ACKNOWLEDGMENT

The author would like to thank Dr. Nan Aye Aye Htwe, Associate Professor, Head of Department of Computer Engineering and Information Technology for her encouragement and kindly advice. The author would like to especially express her deep appreciation to her supervisor, Dr. Moe Moe Aye, Associate Professor, Department of Computer Engineering and Information Technology, Mandalay Technological University for her close supervision, helpful advice, encouragement and invaluable guidance. The author would also thank to her parents, all her friends and all the teachers who taught her throughout the whole life.

REFERENCES

[1] A. K. Jain, "Data clustering: 50 years beyond K-Means," Lecture Notes in Computer Science, Springer, vol. 5211, pp. 3-4, 2008.

[2] L. Kaufman, and P. J. Rousseeuw, "Finding groups in data: An introduction to cluster analysis," John Wiley and Sons, March, 1990.

[3] R. Mawati, I Made Sumertajaya, F. Mochamad Afend "Modified Centroid Selection Method of K-Means Clustering," *IOSR Journal of Mathematics (IOSR-JM, Volume 10, Issue 2 Ver. III (Mar-Apr. 2014), PP 49-53.*

[4] R. Syal, Dr V. Vijaya Kumar, " Innovative Modified K-Mode Clustering Algorithm," International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 4, July-August 2012, pp.390-398.

[5] X. Rui "Survey of clustering algorithms," IEEE Transactions on Neural Network, 2005.

[6] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," KDD Workshop on Text Mining, 2000.

[7] C. Xuhui, X. U. Yong, "K-Means Clustering Algorithm with Refined Initial Center," International Conference on Biomedical Engineering and Informatics, Tianjin, (2010).

[8] D. J. Kalpana, and P.S. Nalwade, "Modified K-Means for Better Initial Cluster Centres," International Journal of Computer Science and Mobile Computing, IJCSMC, Vol.2, Issue.7 July (2013).

[9] O. A. Mohamed Jafar and R. Sivakumar, "Distance Based Hybrid Approach for Cluster Analysis Using Variants of K-Means and Evolutionary Algorithm," Research Journal of Applied Sciences, Engineering and Technology, June (2014).

[10] S. Mubeena, U. B. Ahmedand, U.B. Raheem, "Optimizing Data Using K-Means Clustering Algorithm," International Journal of Engineering Research and Applications, Vol.5, Issue 4 April (2015).

[11] G. Navjot and C. Tejalal, "A High Dimensional Clustering Scheme for Data Classification," International Journal of Engineering Research and Applications, Vol 5, Issue 9, September (2015).

[12] P. Larma, "Clustering System Based on Text Mining Using The K-Means Algorithm," Turku University of Applied Sciences Thesis, Information Technology, December (2013).

[13] S. Singh Raghuwanshi and P. N. Arya, "Comparison of K-Means and Modified K-Means Algorithm for Large Datasets," International Journal of Computing, Communications and Networking, Vol.1, No.3, November- December (2012).

[14] Teknomo, K., "K-Means Clustering Tutorials," (2007).

[15] J. B. Lovins, "Development of a stemming algorithm," Mechanical Translation and computational linguistics, Vol.11, 1968.