Efficient Document Clustering System Based on Probability Distribution of K-Means (PD K-Means) Model

Tin Thu Zar Win[1],
Nang Aye Aye Htwe[2],
Department of Computer Engineering and Information Technology,
Mandalay Technological University, Myanmar
[1]zarzar84mtu@gmail.com
[2]ayehtewnge@gmail.com

and Moe Moe Aye[3]
Department of Information Technology,
Mandalay Technological University, Myanmar
[3]moeaye255@gmail.com

Abstract - **In document clustering system, some documents with the same similarity scores may fall into different clusters instead of same cluster due to calculate similarity distance between pairs of documents based on geometric measurements. To tackle this point, probability distribution of K-Means (PD K-Means) algorithm is proposed. In this system, documents are clustered based on proposed probability distribution equation instead of similarity measure between objects. It can also solve initial centroids problems of K-Means by using Systematic Selection of Initial Centroid (SSIC) approach. So, it not only can generate compact and stable results but also eliminates initial cluster problem of K-Means. According to the experiment, F-measure values increase about 0.28 in 20 NewsGroup dataset, 0.26 in R8 and 0.14 in R52 from Reuter21578 datasets. The evaluations demonstrate that the proposed solution outperforms than original method and can be applied for various standard and unsupervised datasets.**

Keywords - **Initial Centroid, Probability Distribution, PD K-Means, SSIC**

I. INTRODUCTION

In data mining, different techniques are utilized to analyse data using computer based algorithms. These algorithms measure the similarity and dissimilarity among available set of data. Using this evaluation the data patterns are recovered. The analysis of data is performed in both supervised and unsupervised manner. The supervised technique supports the classification techniques and the unsupervised technique supports the clustering techniques for data analysis [1]. The proposed work is devoted to understand and develop an efficient and accurate unsupervised technique which provides efficient results and can resolve the deficiencies of available clustering techniques.

In order to develop such an efficient and accurate clustering algorithm, a number of clustering algorithms such as K-Means clustering, C-means clustering and other techniques are studied. Among them the K-Means algorithm is more effective and frequently used algorithm for clustering high dimensional data. Moreover, some key issues such as fluctuating accuracy, high error rate and running the algorithm in multiple times for better cluster result are observed and targeted to improve K-Means clustering algorithm with stable and more efficient cluster result without

running the algorithm in multiple times. Therefore K-Means algorithm has been investigated in detail and the recovered facts show that K-Means has others weakness such as sensitivity of noise and isolated data points, sensitivity of initial value and unfitting to non-convex cluster [2-3]. In order to resolve the issues in traditional K-Means algorithm, a new PD K-Means clustering algorithm that based on some distribution clustering algorithms is proposed to find the optimum solution.

The most common used of clustering methods is based on geometric distance-based similarity measures in clustering data objects. It is rare in use of distribution regarding the object in traditional clustering algorithms. Clustering distributions has appeared in the area of information retrieval for clustering documents [4]. Topic model is one of the clustering tools which are based on distribution of clustering documents. It is widely used for dimensionality reduction in text collections. Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA) and Expectation Maximization (EM) algorithm which are based on probability distribution of documents were developed in text mining communities. They are generative probabilistic models which can be used to extract topics from text data in the form of word distributions [5].

In this presented work, K-Means algorithm is modified for finding the better and stable performance of clustering. The proposed PD K-Means overcomes the problems facing in the traditional K-Means algorithm by using probability distribution of similar documents instead of geometric measurement. Besides, the improvement is made on the traditional clustering by implementing SSIC approach for initial cluster centroids selection. If the optimum centroids from the data are selected then the issues such as accuracy fluctuation and performance issues are minimized [6]. Therefore the SSIC approach is implemented for selection of initial centroids.

The rest of the paper is organized as follows: Section 2 describes the required background theory of proposed clustering model. In Section 3, it explains the detail procedures of proposed PD K-Means clustering model. And then, experimental results are presented in Section 4. Finally, conclusion is given in Section 5.

II. BACKGROUND THEORY

Clustering of text documents can be a challenging task due to the high dimensionality and the sparsity of features. Clustering is a method of unsupervised classification, where data points are grouped into cluster based on their similarity. The main goal of all clustering algorithm is to produce the efficiency and effectiveness for the proposed result [7]. There are several approaches for clustering process that based on the used techniques, i.e., Distribution based approach, distance-based approach, density-based approach, graph-based approach, supervised and unsupervised learning approach, neural networks and machine learning techniques, etc. [8].

K-Means clustering algorithm is one of the most widely used partition based clustering algorithm that arranges the documents in order such that a document is close to its related document on the basis of similarity measure. Euclidean metric which is the most commonly used distance measures is used in calculation of similarity measure for K-Means. The idea is to classify the data into k clusters where k is the input parameter specified in advance through iterative relocation technique which converges to local minimum. The algorithm proceeds by randomly defining k centroids and assigning a document to the cluster that has the nearest centroid to the document. The centroid initialization plays an important role in determining the cluster assignment in effective ways. It is suited for clustering a large document dataset due to its linear time complexity [9].

Several attempts have been reported to solve the cluster initialization problem. SSIC K-Means algorithm which can produce significantly better and stable clustering solutions is a novel centroid selection algorithm for K-Means.

Selecting initial centroids approach is the main part of K-Means clustering because this can affect the accuracy of clustering. SSIC K-Means algorithm eliminates the random centroid selection approach by replacing systematic centroid selection approach. This approach is helpful in selecting significant centers and improves the cluster quality [6].

Distribution based approach is the method which explore the statistical computation. In this approach, the distribution assumed to fit the dataset and then the objects are evaluated whether those objects are fit or not based on the underlying model of the data. This approach is good but impractical since it needs prior data distribution and the high computation cost [8]. Topic modelling has become one of the most popular probabilistic text modelling techniques and quickly been accepted by machine learning and text mining communities. The most inspiring contribution of topic modelling is that it automatically classifies documents in a collection by a number of topics and represents every document with multiple topics and their corresponding distribution. A word distribution can intuitively represent a topic by assigning high probabilities to words characterizing a topic. A topic model is a generative model for documents; it specifies a probabilistic process by which documents can be generated from a set of topics, where a topic is represented by multinomial distribution of words, i.e. a unigram language model. In the document generation process, one usually first chooses a distribution over topics. Then for each word in the document, one chooses a topic at random according to this distribution and draws a word from that topic. However, such multinomial distribution based methods cannot be applied to general cases where the types of distributions are not multinomial. The two basic representative topic modelling approaches are Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) [10].

EM (Expectation Maximization) algorithm is another clustering tool which offered in statistic. It computes probabilities of cluster memberships based on one or more probability distributions. The most usual clustering algorithm will assign observations to clusters to maximize the distances between clusters. The EM algorithm does not compute actual assignments of observations to clusters, but classification probabilities. In other words, each observation belongs to each cluster with a certain probability. The goal of the clustering algorithm then is to maximize the overall probability or likelihood of the data, given the (final) clusters [11].

III. PROPOSED CLUSTERING MODEL

In this proposed model, the most important contribution is to introduce probability distribution of similar documents instead of geometric measurement. The proposed research derives a unique probability distribution equation which is described in equation 1 based on some previous probability distribution clustering approach. Because of this approach, proposed method overcomes the well-known problem of choosing suitable similarity schemes for clustering process. Moreover, initial centroid selection method is also modified by implementing SSIC approach. Fig. 1, shows the flow chart of the proposed model. In the proposed system, it mainly consists of four main phases: document collection, document pre-processing, initial centroid selection and documents clustering based on probability distribution.

In document collection, documents were collected from two standard datasets such as Reuters-21578 [12] and 20-newsgroups [13]. By using these two standard datasets, the text documents are randomly selected from each category and divided into groups for clustering.

In all clustering algorithms, all documents which are need to pass the pre-processing steps for reducing dimensions. Pre-processing is a very important step since it can affect the result of a clustering algorithm. So it is necessary to pre-process the data sensibly. In this proposed model, two techniques namely, removing stop words and stemming algorithm are used. Removing stop words starts the first process. The stop words are words that carry

no information and meaningless when we use them as search term (keyword). 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.

In initial centroid selection phase, SSIC K-Means approach is used for selecting the accurate initial centroid. It starts by finding the maximum distance objects as initial centroids instead of randomness initial centroid. At first, the similarity matrix of all documents is calculated by using Euclidean distance metric. From these results, the first maximum distance document pair are assigned as first initial centroids. If the number of k cluster is greater than 2, the second maximum distance is needed to choose for third centroid. Then, the average distance is calculated between the selected document pairs and the maximum average distance objects are assigned as the other initial centroid. So, average maximum object is recalculated until the number of k-cluster.

In document clustering phase, remaining documents are grouped depending on the result of initial centroid selection process. One document is randomly drawn from the dataset and computes probabilities of cluster memberships based on proposed probability distributions equation. Then for each word in the document, the relationship between selected document and previous cluster is calculated and clustered similar documents to its corresponding cluster according to this distribution. The probability of each document is needed to calculate for clustering to corresponding cluster. According to probability result, the document which has the overall maximum probability value is assigned to its' related cluster. So, the probability of all process until no more documents is recalculated as an iterative in dataset.

$$P(D_i, C_j) = \sum_{i=1}^{k} P(D_{wi}) * P(w_i \mid C_j) \quad (1)$$

Where, $P(D_{wi}) = \frac{\text{no : of wi count in selected Doc}}{\text{Total word count in selected Doc}}$,

$$P(w_i \mid C_j) = \frac{\text{no : of wi count in cluster}}{\text{Total word count in cluster}}$$

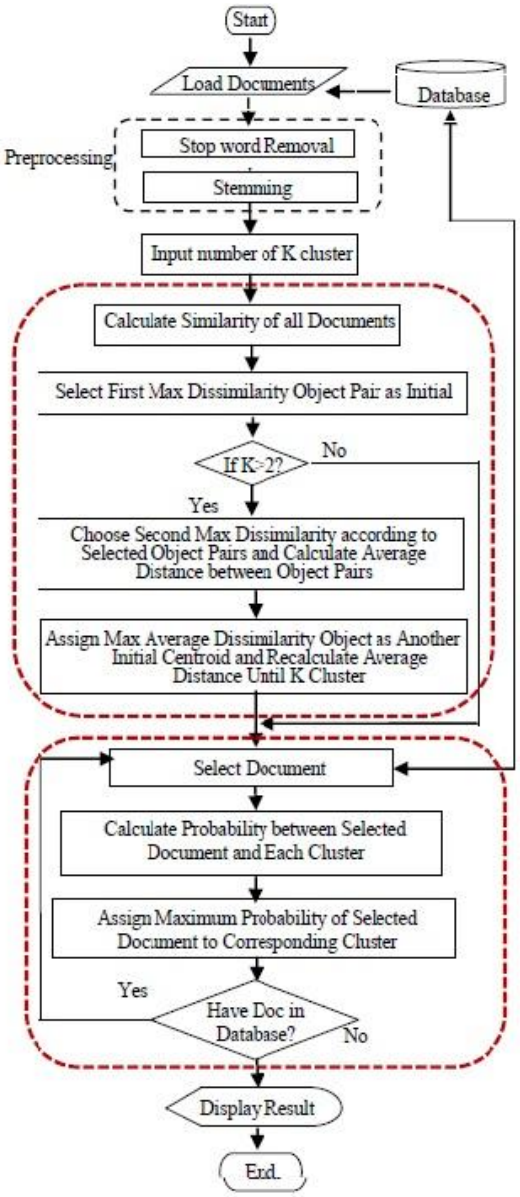

Fig. 1 The Flowchart of PD K-Means Clustering Model

IV. EXPERIMENTAL RESULTS

In this section, two standard datasets, Reuters-21578 [12] and 20-newsgroups [13] were used to experimentally evaluate the accuracy and efficiency of the proposed method.

The second dataset is 20-newsgroups dataset which is a collection of news articles collected from 20 different sources. From 20-newsgroups, 50 documents are randomly selected from each category and developed the data set 20ns consists of 1000 documents. The

two sub-collections, R8 and R52 dataset are chosen from Reuters-21578 dataset for text categorization tasks. The proposed model calculates the precision and recall of each cluster formed and based on their values it determines the F-measure values of the result to obtain the relevancy of the documents in clusters. The results obtained by applying both algorithms to above mentioned datasets are shown in fig. 2.

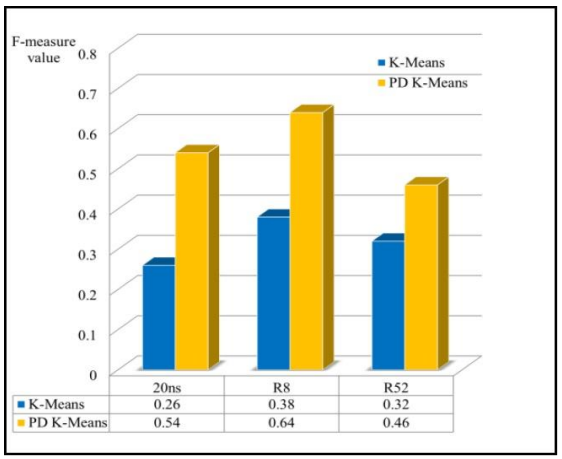

Fig. 2 F-Measure Results of Two Different Algorithms When Run on Three Datasets

It is shown that the F-measure value of 20ns is 0.54, R8 is 0.64 and R52 is 0.46 in proposed PD K-Means algorithm. It is evident that F-measure value is higher for all datasets in PD K-Means algorithm as compared to the original K-Means algorithm. Therefore, proposed algorithm gives better results than the original algorithm.

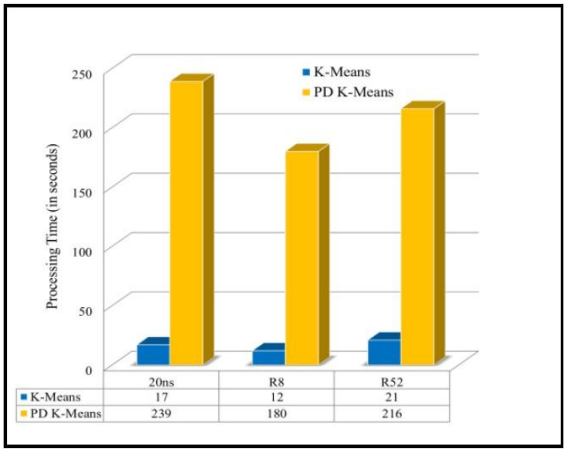

Fig. 3 Comparisons of Processing Time between K-Means and PD K-Means Clustering Algorithm

The processing time of proposed algorithm and K-Means clustering algorithm is given in fig. 3. According to the given diagram, the proposed algorithm consumes more time as compared to the traditional algorithm on these three datasets.

V. CONCLUSIONS

In this article, an effective PD K-Means approach for document clustering system is presented. Because of SSIC approach, it is helpful in selecting significant cluster centers and eliminates initial cluster problem of K-Means. Therefore, it can give stable cluster quality in efficiently. In addition, clustering with probability distribution provides more accurate cluster result. Therefore, PD K-Means algorithm improves cluster quality with reduced complexity and can be applied for various unsupervised and standard datasets. However, designing of perfect clustering algorithm cannot complete yet. The proposed clustering technique is accurate by processing more execution time; the only limitation is that it consumes more time as compared to the traditional clustering algorithm. As a future research work, it is necessary to make further amendments to this clustering algorithm with reduced time consuming on various benchmark datasets. As a second future direction, more evaluation and comparison to other models can help to improve the performance of proposed model.

VI. ACKNOWLEDGEMENT

The author especially would like to take this opportunity to express my sincere gratitude, respect and regards for supervisor Dr. Nang Aye Aye Htwe, Professor, Head of Department of Computer Engineering and Information Technology, Mandalay Technological University. The author would like to express her deep appreciation to teacher, Dr. Moe Moe Aye, Professor, Department of Information Technology, Technological University (Mandalay) for her helpful advice and encouragement. She is also thankful to all her teachers from Department of Computer Engineering and Information Technology, for their help and

encouragement.

REFERENCES

(Arranged in the order of citation in the same fashion as the case of Footnotes.)

[1] Jiawei, H. and Micheline, K. (2006). "Data mining Concepts and Techniques (2nd Ed.)". Data Mining Concepts and Techniques, <http://akademik.maltepe.edu.tr/~kadirerdem/772s>.

[2] Juntao, W. and Xiaolong, S. (2011). "An improved K-Means clustering algorithm". 978-1-61284-486-2/111, IEEE.

[3] Navjot, G. and Tejalal, C. (2015). "A High Dimensional Clustering Scheme for Data Classification". International Journal of Engineering Research and Applications, Vol. 5, No. 9, (Part-1), ISSN: 2248-9622, pp. 101-106.

[4] Bin, J. and et al. (2011). "Clustering Uncertain Data Based on Probability Distribution Similarity". 1041-4347/11, IEEE Transactions on Knowledge and Data Engineering.

[5] Onan, A., Bulut, H., and Korukoglu, S. (2016). "An Improved Ant Algorithm with LDA Based Representation for Text Document Clustering". Journal of Information Science, pp. 1-18.

[6] Win, T.T.Z. and Aye, M.M. (2016). "Systematic Selection of Initial Centroid for K-Means Document Clustering System". The Seventh International Conference on Science and Engineering, Yangon Technological University, Myanmar.

[7] Mehdi, N. and Shima, F.Y. (2012). "A New Cooperative Algorithm Based on PSO and K-Means for Data Clustering". Journal of Computer Science, Vol. 8, No. 2, pp. 188-194, ISSN: 1549-3636.

[8] Djoko, B.S. and et al. (2014). "Rough K-means Outlier Factor Based on Entropy Computation". Research Journal of Applied Sciences, Engineering, and Technology, ISSN: 2040-7459.

[9] Kalpana, D.J. and Nalwade, P.S. (2013). "Modified K-Means for Better Initial Cluster Centers". International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 2, No. 7, pp. 219-223.

[10] Gao, Y., Xu, Y., and Li, Y. (2015). "Pattern-based Topics for Document Modelling in Information Filtering". IEEE Transactions on Knowledge and Data Engineering, Vol. 27, No. 6.

[11] Duk Kim, H. and et al. (2012). "Enriching Text Representation with Frequent Pattern Mining for Probabilistic Topic Modelling". University of Illinois at Urbana-Champaign, ASIST.

[12] <http.//www.daviddlewis.com/resources/testcollections/reuters2158>.

[13] <http.//www.cs.cmu.edu/TextLearning/datasets.html>.