IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-6, June 2018 E-ISSN: 2347-2693

Optimizing Document Clustering for Dimension Reduction using improved k-means

J. Verma[1], N. Verma[2]

[1] Dept. of CSE, Deenbandhu Chotu Ram University of Science & Technology, Murthal, Sonepat, India
[2] Dept. of CSE, Deenbandhu Chotu Ram University of Science & Technology, Murthal, Sonepat, India

Corresponding Author: vermajyoti051@gmail.com Tel: 9050931011

Available online at: www.ijcseonline.org

Accepted: 09/Jun/2018, Published: 30/Jun/2018

***Abstract*—** Clustering is the process for grouping of similar document into a single cluster and dissimilar documents in other clusters. Document clustering is the process of grouping similar text documents in a single cluster. K-means clustering algorithm is a center predictable approach which selects initial centers randomly. In this paper, improved k-means clustering algorithm is used for text documents which predicts centers manually. Standard k-means uses cosine similarity but improved k-means uses Euclidean similarity measures for grouping similar documents in a single cluster. According to experimental results, accuracy of improved k-means is high as compared to existing k-means algorithm. Performance of proposed algorithm is measured in terms of F-measure, Precision, time and recall.

Keywords—Clustering, Document clustering, Tf-Idf, K-means, Euclidean similarity.

I. INTRODUCTION

Clustering is a unsupervised technique of data mining which is helpful for making clusters from the large amount of datasets. With the help of clustering, similar data put in a single cluster and different data put in other cluster [1]. This technique has been used in many problems of data mining like as for building operation among complex datasets, for finding association between objects and for making generalization. This application has been used in many areas like engineering, biomedical, social science, life science, computer science and so on.

Document Clustering is the grouping of similar text documents in a proper cluster. Now-a-days, On the web, documents are increasing in a rapid speed. So, it is important to group similar documents together. Document clustering technique consists various steps. First of all, fetch the dataset from web. After this tokenization and preposition is performed on this dataset. Tokenization provides a set of tokens and preposition provides a set of tokens after stop-word removal, special symbol removal, stemming etc.to vector space model [2]. Vector space model is retrieval process which works as Tf-Idf model. For calculating distance between clusters, similarity measures applied.

The results after experiment shows that the improved k-means algorithm which is advanced algorithm consumes small time for text document clustering as compared to actual k-means algorithm. F-measure factor of improved k-means algorithm is high as compared to actual k-means algorithm.

In this paper, there is brief overview of text document clustering process using improved k-means algorithm. Here, V sections are using. Section I for introducing the paper, Section II for showing the related work of various researchers, Section III for explaining proposed work, Section IV for showing the results after experiment and Section V for concluding this paper.

II. RELATED WORK

Improved k-means clustering comes under partition-based algorithm. It is a method which is used to initialize centroids [3]. Several other clustering algorithm are proposed to cluster the documents including Bisecting K-means Methods [4] which splits the all points sets in 2 clusters, choose one among them and split and repeat method till k clusters are made. Hybrid bisects k-means [5] clustering is used as a combination of bisects k-means and divisive hierarchical algorithm for optimal clusters. Novel algorithm [6]is used for automatic clustering and eliminates the drawbacks of K-Means algorithm.

This paper [7] uses a tree-based document similarity for clustering the documents which extract the phrases and words sequence from documents. An inter passage approach [8] uses k-means algorithm for text documents clustering the

segments on the basis of similarity measure. Genetic clustering algorithm [9] is used to deal with clustering aggregation problem.

III. METHODOLOGY

K-means clustering algorithm is good for small set of datasets but when the size of dataset increases k-means algorithm doesn't cluster the documents as good. The results of clusters formed depends on initial centroids values which are selected randomly and it work well with global clusters only. [10] k-means is based on selection of predefined clusters only. To cope up with these issues, we started to cluster the documents with improved k-means clustering algorithm for improving F-measure's value.

The procedure of doing work is explained below in figure 1:

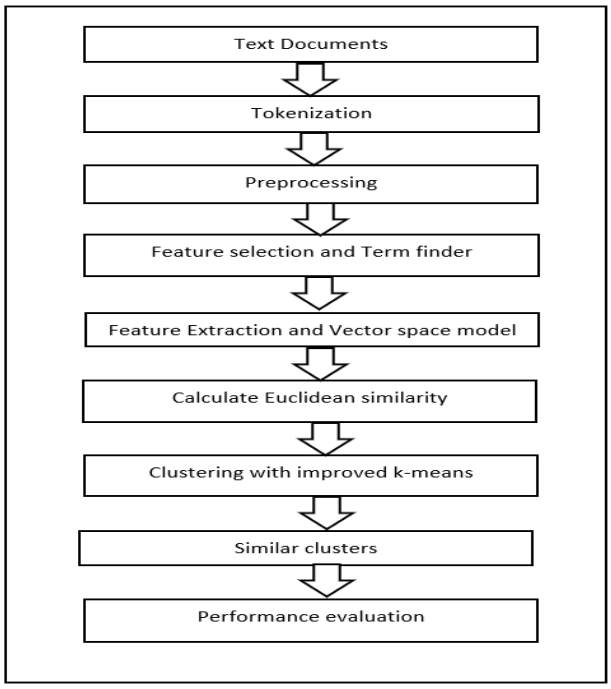


Fig.1 Document Clustering Process

Text Documents: In this step, read the dataset on which document clustering will performed.

Tokenization: In tokenization, splitting of tokens and strings into individual words and tokens.

Pre-processing: Pre-processing is performed on text documents which are taken after tokenization and it generates a set of tokens as output to VSM. This technique provides optimal quality of clusters. Main steps of pre-processing are as follows:

1. Filtering: for removing punctuation marks and special characters '!@#$%^&*():"{}'. This process is also called as special symbol removal.
2. Stop word removal: This process is used for removing stop words. Stop words are the words which have no meaning like example, again, after, about etc.
3. Stemming: In this process, creating base form of every token. Example: base form of tokens dependency, dependent, dependable is depend.
4. Pruning: This process is used for removing low frequency words. Pruning improves predictive accuracy.

Feature Selection: Process of selecting features. Features means variables, attributes, attribute subsets. It selects the relevant features subsets for constructing a model. This technique is used for simplification of a model, for decreasing time complexity.

Term finder: Term finder selects/finds exclusive terms from each available category. The proposed work assigns a threshold value as a weight to each term. If the term frequency is greater than threshold than value is added, else rejected.

For all words in term sets

If (tf (i) > threshold)

Add to set

End

Feature extraction and Vector Space Model:

"For extracting set of keywords from text document, Feature extraction process is used. Vector Space Model(VSM) is a retrieval technique in data mining and is also define as Term frequency Inverse document frequency model. For representing text, this standard algebraic model is used by using feature vector, every text document is express as an n-dimensional trajectory. Importance of corresponding features in the text documents is reflected by the rate of every element in the trajectory. In this model, By determining the distance between document vector, similarity between text documents can be determined. If same keywords present in any text document then they are similar. Term frequency Tf(I,j), is the any ith term's total occurring time in a document. [11] Compared with tf and Boolean feature selection scheme, results show that tf-idf is best for producing clusters. Term frequency is normalized with relevance the greatest frequency of all terms occurring in a text document.

$$Tf(i, j) = freq(i, j) / (max \{f(x,j):w€j\})$$

Where,

i is any term
j is document
x is the term with greatest frequency

Similarly, Text document frequency of any term is calculated as the total documents in which ith term present.

It measured as,

$$Idf (i , j) = \log (D/dfi)$$

Euclidean Similarity: This is most commonly used similarity measure. Euclidean similarity measure is used for measuring the closeness between two text documents. After measuring harmony or closeness between two text documents, clustering will be performed. Euclidean distance is measured as:

$$S=S+ ((a (t)-b (k))^{*1/2})$$

Improved k-means

Suppose a vector of documents $[x_1,x_2,.....x_n]$ is given. Improved k-means clustering algorithm will divide the n documents into k clusters in a way that Euclidean distance between documents is small. It initially predicts centre manually and them perform k-means on datasets of text documents.

Performance Matrix

"Performance evaluation in clustering is measured in terms of F-measure and time. Time is measured as how much time consumes in clustering. F-measure is used for comparing closeness between two clusters. It is a harmonic function which is aggregate of Precision (P) and Recall (R) measure. It is set for measuring accuracy. It is given by:

$$F\text{-measure} = (2*PR)/ (P + R)$$

Precision(P) is determined as the number of true positive(T_P) over the sum of number of true positive and number of false positive(F_P).

$$P = (T_P)/(T_P + F_P)$$

Recall (R) is determined as the number of true positive(T_P) over the sum of number of true positive and the number of false negative(F_N).

$$R = (T_P)/(T_P + F_N)$$

The Proposed Algorithm

"The work is performed on a mini_newsgroups dataset. We proceed with the following algorithm.

Algorithm: Improved k-means

Input: $x = \{x_1, x_2, x_3, ,……, x_n \}$

// set of n data objects.

K // cluster's number.

Output: Set of K clusters.

Phase 1. By using algorithm 2, determining the initial centres of every cluster.

Phase 2. Assigning the data points to the clusters which have shortest distance from these.

Improved k-means algorithm works on two phases. Phase 1 determines the initial centres of every cluster. This phase is used for improving accuracy. Phase 2 used for assigning the data points to the clusters. These data points assign by determining the Euclidean distance between them.

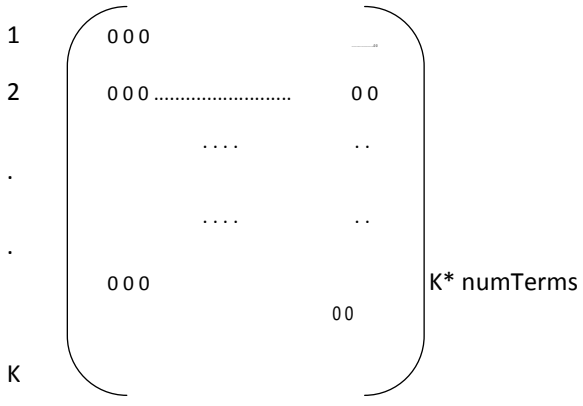
Phase 1: Initial Centre Prediction

Input: VSM, K, n, terms, w

Output: centres

Steps:

1. Create initial centre matrix for clusters and set default to zero.



3. For ki=1 to k

Centers(Ki;t:t+n-1) = w t=t+n

End

Phase 2: Grouping the data points into clusters.

Input: $X=\{X_1,X_2.....X_i.....X_n\}$

$C=\{c_1,c_2.....c_k\}$

Output: Set of K clusters.

Steps:

1. Determine the Euclidean distance between all data points in X and all centroids.
2. For each data point in X, find the closest centroid and assign the cluster.
3. For each data point Xi

3.1 Determine its distance from centroid to current closet cluster

If

Distance <= present nearest distance

Then

Cluster remains same

Set clusteredID (i) = j.

Set near_dist = d (Xi, cj).

Else

For every centroid, compute the distance and assign cluster to nearest centroid.

Set clusteredID (i) = j.

Set near_dist = d (Xi, cj).

End

End

4. Recalculate the centroids until the results of all iterations remain same.

IV. RESULTS AND DISCUSSION

The algorithm is tested on mini_Newsgroup datasets. The work is implemented on complete system including all models discussed in section III in MATLAB. Following three categories were taken.

Table 1 for Categories of dataset:

S. No.	Categories	Symbols
1	Atheism	Alt.atheism
2	Computer graphics`	Comp.graphics
3	Computer operating system-window misc	Comp.os.ms-window.misc

For analysing the result, both existing K-means algorithm and improved K-means algorithm applied on different categories of the dataset. From these three categories, we have 300 text documents of mini_Newsgroup. For analysing results, we applied both algorithm ten times on this dataset. After this, we found that existing K-means clustering algorithm gives distinct results when the centroids are selected randomly but in improved k-means clustering algorithm gives same value in every execution as the centres are predicted manually.

Table 2. and Table 3. shows the value of precision, recall, F-measure and time. In the both tables, we check the values of all factors in 10 attempts. Table 2. Shows the values of existing k-means clustering algorithm in which cosine similarity function is used. This Shows that when centroids are predicted randomly then the results will different in every attempt. Table 3. Shows the value improved k-means clustering algorithm in which Euclidean similarity function is used. This shows that when we predict the centroids manually then the results will same in every attempt. By using improved k-means algorithm, we reduce the time complexity.

Table 2. Accuracy measure of k-means cosine

Kmeans-cosine				
Attempt	Precision	Recall	F-Measure	Time(sec)
1	0.4208	0.3667	0.3166	0.0148
2	0.7585	0.65	**0.6287**	0.2714
3	0.1336	0.3167	0.1879	0.0262
4	0.3643	0.2267	**0.2212**	0.0163
5	0.4651	0.3567	0.2325	0.0129
6	0.7763	0.59	**0.5776**	0.0282
7	0.4744	0.37	0.3418	0.0147
8	0.7743	0.7233	**0.6864**	0.0301
9	0.8036	0.6567	0.6576	0.0148
10	0.4765	0.4633	**0.3814**	0.0189
	0.54474	**0.47201**	**0.42317**	**0.04483**

Table 3. Accuracy measure of I k-means

iKmeans				
Attempt	Precision	Recall	F-Measure	Time(sec)
1	0.8177	0.82	0.818	0.0104
2	0.8177	0.82	0.818	0.0175
3	0.8177	0.82	0.818	0.0082
4	0.8177	0.82	0.818	0.0172
5	0.8177	0.82	0.818	0.0224
6	0.8177	0.82	0.818	0.0084
7	0.8177	0.82	0.818	0.0172
8	0.8177	0.82	0.818	0.0224
9	0.8177	0.82	0.818	0.0085
10	0.8177	0.82	0.818	0.0101
	0.8177	0.82	0.818	0.01423

Table 4. represents that the values of precision, recall, F-measure and time for both existing K-means and advanced K-means clustering algorithm. From the table 4, we found that value of precision, recall and f-measure is high in case of improved k-means (I k-means) clustering algorithm as compared to existing k-means. Improved k-means takes less time in comparison of existing one. So, we can say that the accuracy of improved k-means is high in comparison of existing.

Table 4. values of accuracy for existing and proposed algorithms

	Precision	Recall	F-Measure	Time(sec)
Kmeans-cosine	0.54474	0.47201	0.42317	0.04483
iKmeans	0.8177	0.82	0.818	0.01423

Figure 2. shows the accuracy of proposed algorithm is same in every attempt but the accuracy of existing algorithm vary with every attempt.

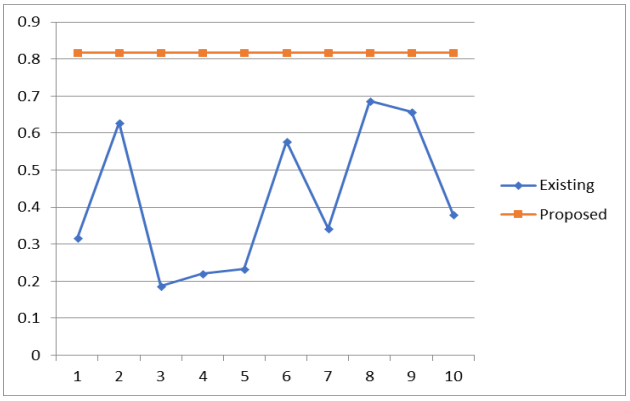


Figure 2. Accuracy comparison of k-means/i k-means

Figure 3 represents the comparison of both the algorithms with respect to time. In this, we found that existing clustering algorithm takes more time as compared to improved k-means.

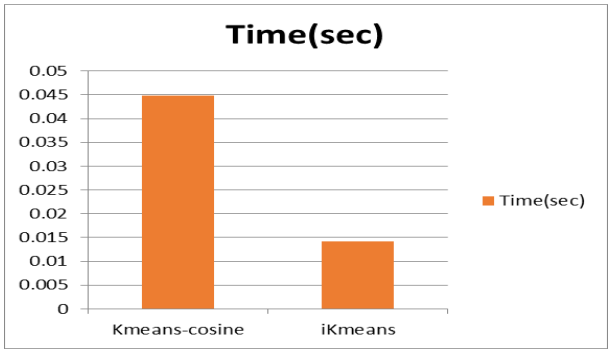


Figure 3. Time comparison for mini_newsgroup

Figure 4 compares the existing k-means algorithm with the proposed k-means algorithm. From this comparison, we found that the proposed k-means algorithm is better than the existing k-means algorithm. In this, accuracy is measured in terms of precision, recall and F-measure. Proposed algorithm shows better accuracy because its values are high.

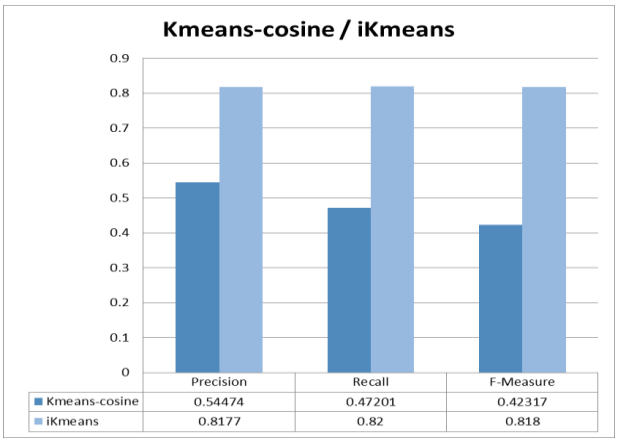


Figure 4. Accuracy measure

V. CONCLUSION AND FUTURE SCOPE

Text document clustering process is used for grouping of similar text documents in a single cluster and dissimilar text documents in other clusters. For performing text documents clustering, k-means clustering algorithm is used. K-means algorithm is a centre predictable algorithm. In existing k-means clustering algorithm, by using cosine similarity measure accuracy is achieved in form of precision, recall, F-measure and time. But in proposed k-means or improved k-means, Euclidean similarity measure is used. By using this, accuracy is high as compared to existing one. In improved k-

means, centres are predicted manually due to this result is same in every attempt and time consumes less. On the basis of above parameters, we can say that improved k-means is better than existing k-means. In future, better performance can be achieved by using other similarity measures like manhattan distance, Minkowski distance, Jaccard distance measures etc. .

REFERENCES

[1] Svadas, T., & Jha, J. (2015). Document Cluster Mining on Text Documents.

[2] Thomas, A. M., & Resmipriya, M. G. (2016). An efficient text classification scheme using clustering. *Procedia Technology*, *24*, 1220-1225.

[3] Punitha, S. C., Jayasree, R., & Punithavalli, M. (2013, January). Partition document clustering using ontology approach. In *Computer Communication and Informatics (ICCCI), 2013 International Conference on* (pp. 1-5). IEEE.

[4] Rai, P., & Singh, S. (2010). A survey of clustering techniques. *International Journal of Computer Applications*, *7*(12), 1-5.

[5] Murugesan, K., & Zhang, J. (2011, July). Hybrid bisect K-means clustering algorithm. In *Business Computing and Global Informatization (BCGIN), 2011 International Conference on* (pp. 216-219). IEEE.

[6] Agrawal, R., & Phatak, M. (2012). Document clustering algorithm using modified k-means.

[7] Rafi, M., Maujood, M., Fazal, M. M., & Ali, S. M. (2010, June). A comparison of two suffix tree-based document clustering algorithms. In *Information and Emerging Technologies (ICIET), 2010 International Conference on* (pp. 1-5). IEEE.

[8] Mishra, R. K., Saini, K., & Bagri, S. (2015, May). Text document clustering on the basis of inter passage approach by using k-means. In *Computing, Communication & Automation (ICCCA), 2015 International Conference on* (pp. 110-113). IEEE.

[9] Zhang, Z., Cheng, H., Zhang, S., Chen, W., & Fang, Q. (2008, June). Clustering aggregation based on genetic algorithm for documents clustering. In *Evolutionary Computation, 2008. CEC 2008.(IEEE World Congress on Computational Intelligence). IEEE Congress on* (pp. 3156-3161). IEEE.

[10] Wang, J., & Su, X. (2011, May). An improved K-Means clustering algorithm. In *Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on* (pp. 44-46). IEEE.

[11] Singh, V. K., Tiwari, N., & Garg, S. (2011, October). Document clustering using k-means, heuristic k-means and fuzzy c-means. In *Computational Intelligence and Communication Networks (CICN), 2011 International Conference on* (pp. 297-301). IEEE.

[12] Sahu, L., & Mohan, B. R. (2014, December). An improved K-means algorithm using modified cosine distance measure for document clustering using Mahout with Hadoop. In *Industrial and Information Systems (ICIIS), 2014 9th International Conference on* (pp. 1-5). IEEE.

Authors Profile

Neetu Verma is currently working as Assistant Professor in the Department of Computer Science & Engineering at Deenbandhu Chhotu Ram University of Science and Technology, Murthal, Sonipat. She has more than 10 years of teaching & research experience in all. She has got Bachelors of Engineering & Masters of Engineering, both, from MDU Rohtak. She is in advanced state of her Ph.D from DCRUST Murthal. She has successfully guided numerous M.Tech scholars. She has published many research papers in reputed Journals & Conferences. Her area of expertise is in Wireless Sensor Network, Compiler design & Operating System.

Jyoti Verma currently completed her Master of Technology from DCRUST,Murthal & Bacchelors of Technology from MDU, Rohtak. She has published her papers in reputed Journals. Her area of interest is in Data Mining.

