International Journal of Engineering Research and General Science Volume 4, Issue 3, May-June, 2016
ISSN 2091-2730

Document Clustering Using Improved K-Means Algorithm

Shreyata khatri[1],Dr. Kanwal Garg[2]
Research scholar,DCSA, Kurukshetra university,kurukshetra
Assistant professor, DCSA Kurukshetra University, kurukshetra

Abstract: Clustering is the process where similar documents are grouped under a single cluster. K-means clustering is a common approach based on selecting initial centroids randomly. In this paper, improved k-means clustering algorithm is used for document clustering by predicting centres manually. The algorithm uses Euclidean similarity measures to place similar documents in proper clusters. Experimental results showed that accuracy of proposed algorithm is high compare to existing algorithm in terms of F-Measure and time complexity.

Keywords: Document Clustering, k-means, Cosine Similarity, Tf-Idf.

I. INTRODUCTION

Nowadays, Documents on web are increasing with a huge speed. Similar documents are required in various purposes like statics, marketing, engineering, medical and other social science. So it is important to group similar documents together. For this purpose document clustering is used. Clustering is the process where similar objects are grouped under a cluster and dissimilar objects under another cluster.

Document clustering process consists of various steps. First off all pre-processing is performed on datasets which provide e set of tokens to vector space model (VSM). [1] VSM is a retrieval process which works on Tf-Idf model. Similarity measures are used for calculating the distance between various clusters.

The Experimental results showed that the proposed algorithm takes less time for clustering compare to existing K-means algorithm and the F-measure score is also very high than existing algorithm.

In this paper there is brief overview of the document clustering process using improved k-means. Section 1 gives the introduction, section 2 explores related work of various researchers, section 3 presents the proposed work, section 4 show experimental results, section 5 concludes the paper.

II. RELATED WORK

Improved k-means clustering comes under partition based algorithm. It is a method which is used to initialize centroids [2]. Several other clustering algorithm are proposed to cluster the documents including Bisecting K Means Methods [3] which splits the set of all points into two clusters, select one of them and split and repeat process until k clusters have been produced. Hybrid bisects k-means [7] clustering is used as a combination of bisects k-means and divisive hierarchical algorithm for optimal clusters. Novel algorithm [8] is used for automatic clustering and eliminates the drawbacks of K-Means algorithm.

This paper [4] uses a tree based document similarity for clustering the documents which extract the phrases and words sequence from documents. An inter passage approach [5] with k-means is used for clustering the segments based on similarity. Genetic clustering algorithm [6] is used to deal with clustering aggregation problem.

III. PROPOSED WORK

K-means algorithm work well for certain documents. As the number of documents increases, k-means algorithm doesn't cluster the documents well. 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 algorithm for improving the value of F-measure.

The flow of our existing work is shown in Figure1. The process consists of various steps such as:

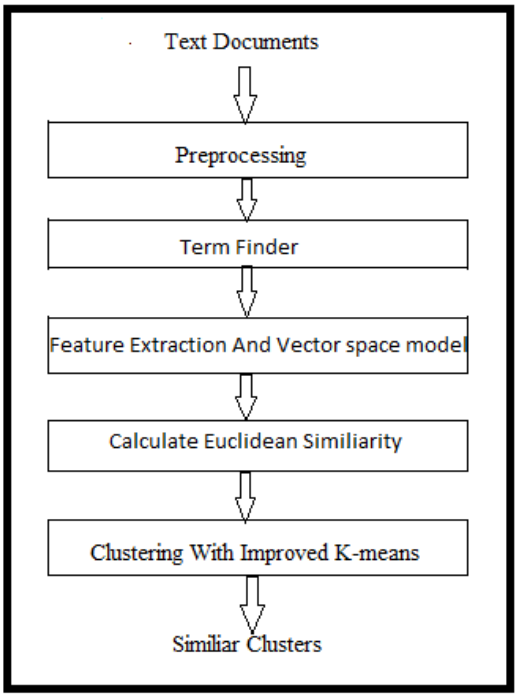


Figure1. Document Clustering Process

1. PREPROCESSING

Pre-processing is performed on plain text documents 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.*
2. *Tokenization: for splitting tokens into individual words and tokens.*
3. *Stop word removal: words with no meanings are removed.*
4. *Stemming: base form of words is formed.*
5. *Pruning: for removal of low frequency words.*

2. *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

3. *Feature extraction and Vector Space Model*

Feature extraction is used to extract a set of keywords from documents. VSM is a retrieval technique in data mining and is also known as Term Frequency Inverse Document Frequency model i.e. TF-IDF model. It is the standard algebraic model for representing text. Each document is represented as an n-dimensional vector using the feature vector. The value of each element in the vector reflects the importance of the corresponding feature in the document. With this model the similarity between documents can be

measured by calculating the distance between document vectors. If the Documents contain the same keywords they are similar. Term frequency tf (i, j), is the number of times a term i occurs in a document. [9]Compared with tf and Boolean feature selection scheme, results show that tf-idf is best for producing clusters. The term frequency is normalized with respect to the maximal frequency of all terms occurring in a document.

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

Where,

i=term in document j.

X=any term with maximum frequency.

Similarly document frequency of a term is the number of documents in which term i occurs.

It is calculated as,

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

Euclidean similarity is the most commonly used similarity measure for calculating the similarity between two documents. It is calculated as:

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

4. Performance Matrix

Performance evaluation in clustering is measured in terms of F-measure. F-measure is used to compare how similar two clusters are. It is a combination of Precision (P) and Recall measure. It is given by:

F-measure= (2*PR)/ (P+R)

Precision (P) is defined as the number of true positives (T_P) over the number of true positives plus the number of false positives (F_P).

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

Recall (R) is defined as the number of true positives (T_P) over the number of true positives plus the number of false negatives (F_N).

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

5. Improved k-means

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

6. The Proposed Algorithm

The work is performed on a mini_newsgroups dataset. we proceeds with the following algorithm.

Algorithm: Improved k-means

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

// set of n data points

K // Number of clusters.

OUTPUT: Set of K clusters.

Phase1. Determine the initial centroids of clusters using algorithm2.

Phase2. Assign each data point to the nearest clusters.

The algorithm works in two phases. First phase calculates the initial centroids for improving accuracy and second phase assign data points to clusters by calculating Euclidean distance between them.

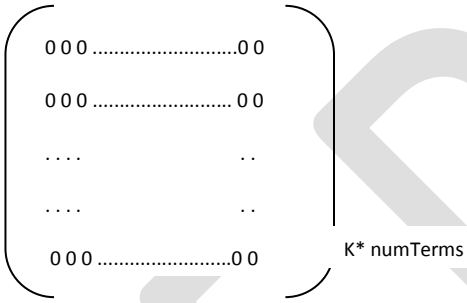
Algorithm2: Initial Center Prediction

Input: VSM, K, n, terms, w

Output: centres

Steps:

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



2. ,
3. For ki=1 to k
   Center (ki: t: t+n-1) =w
   t =t + n

End

Algorithm 3: Assign Data Points to Clusters.

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

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

Output: set of clusters.

Steps:

1. Compute the Euclidean distance for each data point in X to 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. Compute its distance from centroid to present nearest cluster
   If

Distance <= present nearest distance

Then

Cluster remain 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 remain same.

IV. EXPERIMENTAL RESULT

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

Table1. mini_newsgroup

Categories
Alt.atheism
Comp.graphics
Comp.os.ms-window.misc

For analysis of our result, we have applied k-means and improved k-means algorithm on 300 documents from mini_Newsgroup. The k-means clustering algorithm provides different result as the centroids are selected randomly where as in improved k-means value is same for every execution as the centres are predicted manually.

Figure2. Shows the value of precision, recall and f-measure for both existing and proposed algorithm. F-measure has greater value for proposed algorithm as compared to existing algorithm. Also value of precision and recall is better for proposed algorithm compared with existing algorithm.

Table2: Value of Accuracy for Existing and Proposed Algorithm

	Precision	Recall	f-measure
k-means	0.54474	0.47201	0.42317
ik-means	0.8177	0.82	0.818

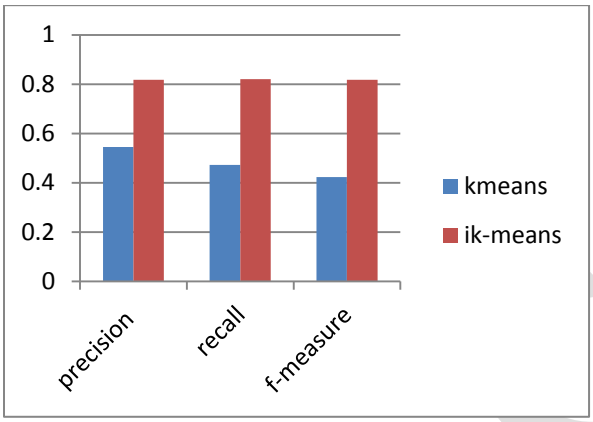


Figure2.accuracy measure

Figure3. Shows the comparison of both the algorithm with respect to time. Existing algorithm takes more time then proposed algorithm.

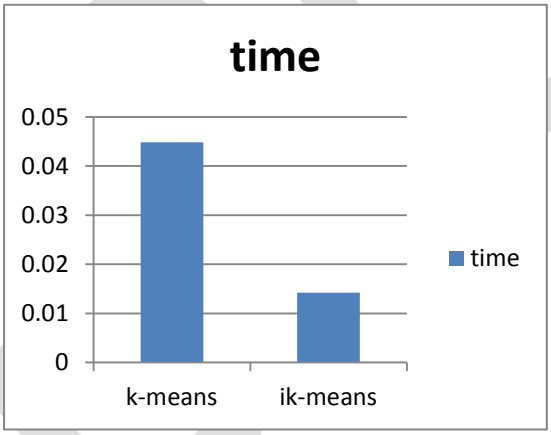


figure3. Time comparison for mini_newsgroup

V. CONCLUSION

Documents clustering are widely used to segregate the similar documents together and dissimilar together. Traditional k-means algorithm work well with certain documents and centroids are selected randomly. With proposed algorithm centroids are predicted manually. Every time results are same. The experimental results have shown that the improved k-means performs better than existing algorithm in terms of accuracy, f-measure and time.

REFERENCES:

[1]Promod Bide, Rajashree Shedge, "Improved Document Clustering Using K-Means Algorithm", International Conference on Electrical, Computer and Communication Technologies (ICECCT- 2015), pp: 1-5, IEEE, 2015.

[2] S.C. Punitha, R. Jayasree And Dr. M. Punithavalli, "Partition Document Clustering Using Ontology Approach", International Conference on Computer Communication and Informatics (ICCCI -2013), Jan. 04-06,pp: 1-5, 2013.

[3]Pradeep Rai. Shubha Singh," A Survey Of Clustering Techniques", International Journal Of Computer Applications, volume 7,pp:1-5, 2010.

[4] Mohammad Rafi, Mehdi Moujood, Murtaza Munawar Fazal, Syed Muhammed Ali, " A Comparsion Of Two Suffix Tree Based Document Clustering Algorithms", International Conference On Information And Emerging Technologies (ICIET- 2010), June 14-16, pp:1-5, IEEE, 2010.

[5]Rupesh Kumar Mishra, Knika Sain, Sakshi Bagri, "Text Document Clustering On The Basis Of Inter Passage Approach By Using K-Means", International Conference On Computing, Communication And Automation,(ICCCA- 2015), may 15-16, pp:110-113,IEEE, 2015.

[6] Zhinya Zhanga,Hongmes Cheng, Shuguang Zhang, Wanli Chen, "Clustering Aggregation Based On Genetic Algorithm For Document Clustering", IEEE congress on evolutionary computation, june 1-6, pp: 3158-3161, IEEE, 2008.

[7]Keerthiram Murugesan ,Jun Zhang, "Hybrid Bisect K-Means Clustering Algorithm", International Conference on Business Computing and Global Informatization , july 29-31,pp:216-219,IEEE, 2011.

[8] Ranjana Agrawal, Madhura Phatak, "Document Clustering Algorithm Using Modified K-Means ", fourth international conference on advances in recent technologies in communication and computing(artcom-2012),pp: 294-296,IEEE, 2012.

[9] Vivek Kumar Singh, Nisha Tiwari,Shekhar Garg, "Document Clustering Using K-Means,Heuristic K-Means And Fuzzy C-Means", International Conference On Computational Intelligence And Communication On Networks (CICN-2011), oct 7-9, pp:297-301, IEEE, 2011.

[10] Junto Wang, Xialong Su, "An Improved K-Means Clustering Algorithm", pp: 44-46,IEEE,2016.

[11]Lokesh Sahu,Biju R.Mohan, "Improved K-Meansclustering Using Modified Cosine Distance Measure For Document Clustering Using Mahout With Hadoop", 9th International Conference On Industrial And Information System(ICIIS-2014), Dec 15-17, pp: 1-5, IEEE, 2015