Volume 3, Issue 5, May 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Customized Vector Space Model Implementation in Document Clustering to Enhance the Performance

M. Praveen*
M .Tech 2nd year, Dept of CSE,
ASCET, GUDUR, India

Dora Babu Sudarsa
Associate Professor , Dept of CSE
ASCET, GUDUR, India

Abstract— Document clustering is the task of grouping a set of documents into clusters so that the documents in the same cluster are similar to each other than to those in other clusters. One of the applications of document clustering is in web search engine retrieval system to help the users find relevant information quicker, and allow them to focus their search in the appropriate direction. Kmeans is a commonly used algorithm for document clustering, but it has some disadvantages. The main limitations of K-means are: 1) The number of clusters K has to be given as input and 2) Based on the initializations it converges to different local minima. 3) It is slow and cannot be used for large number of data novel algorithm to eliminate all these basic drawbacks of K-means.

Keywords— Document clustering, K-means, Cosine similarity, Threshold

I. INTRODUCTION

K-means clustering is one of the popularly used techniques for document clustering. In our research, we have implemented K-means algorithm and found its followinglimitations: 1) The number of clusters K has to be given as input 2) Based on the initializations it converges to different local minima. 3) It is slow and cannot be used for large number of data points.4) It cannot handle empty clusters.To overcome all these drawbacks of K-means, we have modified VSM input to Kmeans in various ways as shown in table 3 and solved last three issues. Then we developed a new algorithm using Cosine Similarity Threshold which resolves all the issues. Comparison is done using 20Newsgroup-18828 text dataset.

II. Related Work

K-means clustering is a partitioning type clustering algorithm. In this method, data is partitioned into K clusters. The clusters are identified by a set of points called the cluster centers. The data points belong to the cluster whose center is closest. There are various algorithms developed to improve efficiency of K-means algorithm. Initial cluster centers for K-Means clustering are computed based on an efficient technique for estimating the modes of a distribution [1]. K-means algorithm is modified which uses Jaccard distance measure for computing the most dissimilar k documents as centroids for K clusters [2]. Unsupervised feature selection method is used to reduce the dimension of document feature space and then a novel partitioning based algorithm is used which select initial cluster centroids in the process of clustering by the size and density of cluster in the datasets [3]. Comparison of Euclidean K-means(K-means), Spherical K-means (SK-means) and unsupervised Principal Direction Divisive Partitioning (PDDP) algorithms is done [4].In all these papers though efficiency of K-means is improved, major drawback of K-means that is “input number of clusters K” still remains.

III. System Architecture

Our document clustering technique uses vector space model. The stages of document clustering are as explained below:
System takes text document dataset as input. Preprocessing is done on all documents to remove some non informative words .e.g. “in, and, the…”.Feature vector is extracted from the dataset by taking keywords with maximum frequency.
Vector space model (VSM) using tf-idf formula computes the term frequency and weight of each word in the documents. Final vector space model is a numerical matrix representation of text data. It is then normalized. Now K-means algorithm using Cosine distance is applied to this vector space model. There were some implementation issues: 1) the document which did not contain a single word from the keyword list gets the cosine value ∞ and it cannot be clustered .2) If we give similar documents to K-means with clusters K>1, it clusters those documents in different clusters. We modified VSM input to K-means in various ways and solved first issue. Then we developed a new algorithm using Cosine similarity Threshold which resolves both the issues.

IV. Algorithm Implementation

As mentioned before, K-means algorithm worked well for a few documents with maximum keywords. But when the number of documents is increased, it could not cluster the documents because some documents did not have words matching in the feature vector which makes their VSM value zero and cosine distance ∞.To overcome this problem we tried modifying input to K-means which is VSM in various ways. It is as shown in table 1.First we calculated Cosine similarity (the cosine of the angle of two vectors) for the above VSM (Cosine VSM). It is a square matrix of size Number

of Documents x Number of Documents. We reduced the columns to 9 by taking various parameters as mentioned in table 1.

Table 1. Modified Input to K-means

Sr. No.	Algorithm using Cosine VSM	Algorithm using VSM	Algorithm using Cosine VSM nonzero values	Algorithm using Cosine VSM
1	Document itself	Document itself	Document itself	Document itself
2	Document with which it has maximum cosine similarity relation	Document with which it has maximum cosine similarity relation	Document with which it has maximum cosine similarity relation	Document with which it has maximum cosine similarity relation
3	Maximum Cosine similarity value	Maximum Cosine similarity value	Maximum Cosine similarity value	Maximum Cosine similarity value
4	Median	Median	Median	-
5	Mean	Mean	Mean	-
6	Variance	Variance	Variance	-
7	Standard deviation	Standard deviation	Standard deviation	-
8	Covariance	Covariance	Covariance	-
9	Entropy	Entropy	Entropy	-

These modified inputs not only overcome the limitation of zero document clustering but also give better clusters compared to K-means with improved time complexityResults are shown in figure 2 and 3.From these results we observe that clustering is improved using Cosine distance measure than only VSM. To resolve the issue of "manually input K", we started with Cosine similarity matrix. The documents above a threshold value (chosen from the Cosine similarity matrix) are clustered into C clusters. Then non clustered documents are either grouped with these C clusters or given next cluster numbers depending on their cosine similarity. Thus C clusters are automatically generated. The document with zero VSM value will not be clustered and it is separated out ..The reason is its cosine value is below threshold value. Rests of the documents are clustered successfully.

V. Experimental Results

20Newsgroup text dataset contains 18828 documents of 20 categories. They can also be clustered in 6 main categories. After preprocessing, Vector space Model (VSM), Normalized VSM and Cosine VSM are computed.1) K-means and modified input algorithms using K-means as listed in table 1 are executed. The results are as shown in Figure 1 and 2 show considerable improvements in F-measure as well as in time taken to execute these algorithms.

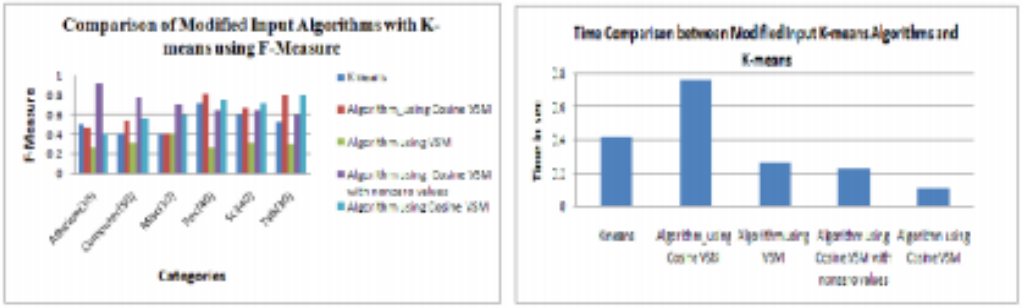

Figure 1and 2: Comparison of Modified Input Algorithms using F-Measure (Left) and Time comparison of Modified Input Algorithms (Right)

2) K-means and our proposed algorithm using cosine similarity threshold are applied on different keyword sets as well as different number of documents set. We had following observations:a) K-means worked only for the Vector space model with 200 and 500 documents but because of zero clustering error it did not work for large no. of documents set. Our algorithm resolves this problem and it works for any document set with any number of features. The zero clustering is shown in figure 3. As shown, document number 66 cannot be clustered and it is shown in zero clusters.

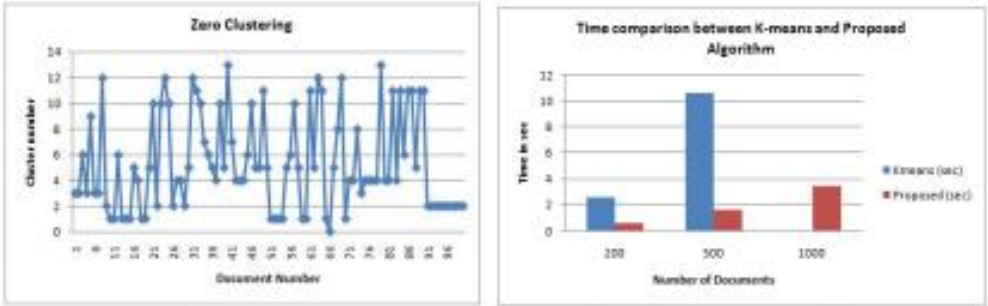

Figure 3and 4: Zero Clustering (left) and Time comparison of Proposed Algorithm (right)

b) As the number of documents increases K-means takes much more time than our algorithm. For e.g. for 500 documents K-means takes 10.63 seconds in MATLAB whereas our algorithm takes 1.52 seconds as shown in figure 4. K-means did not work for documents > 500 showing time zero seconds whereas our algorithm takes comparatively less time

VI .Conclusion And Future Work

Document Clustering for web search engine retrieval system can be done efficiently using different clustering algorithms. K-means is a basic method used for it since it is easy to implement and understand. K-means has some serious drawbacks which are overcome by modifying input to K-means in various ways as shown in the experimental results. We are also working on how to generate the cluster number K automatically and optimize the clusters irrespective of K using new document clustering algorithm using cosine similarity threshold .This algorithm overcomes the drawback of zero clustering effectively. The results of automatically generated K we will be presenting soon.

REFERENCES

[1] Xiaoping Qing, Shijue Zheng "A new method for initializing the K-means clustering algorithm" 2009 Second International Symposium on Knowledge Acquisition and Modeling

[2] Mushfeq-Us-Saleheen Shameem, Raihana Ferdous, "An efficient K-means Algorithm integrated with Jaccard Distance Measure for Document Clustering" 2009 IEEE

[3] Zonghu Wang, Zhijing Liu, Donghui Chen, Kai Tang "A New Partitioning Based AlgorithmFor Document Clustering" 2011 Eighth International Conference on Fuzzy Systems and

[4] Knowledge Discovery (FSKD). V. Mary Amala Bai,Dr. D. Manimegalai "An Analysis of Document Clustering Algorithms"

[5] 2010 IEEE . Michael Steinbach George Karypis Vipin Kumar "A Comparison of Document Clustering Techniques

