A Divide and Conquer strategy for Document Clustering using Reformed K-Means

V.R. Mounika [1], Dr.B. Prajna [2]

[1]Department of Computer Science and Systems Engineering,
Andhra University, Visakhapatnam, India.
raghamounikavanapalli@gmail.com

[2]Professor,
Department of Computer Science and Systems Engineering,
Andhra University, Visakhapatnam, India.
prajna.mail@gmail.com

Abstract— Clustering documents based on the topic is an interesting challenge. The actual challenge is that a text document comprises of various topics, and hence it is hard to put it in a cluster representing one concrete topic. In this paper an attempt to cluster the documents having variety of topics based on the divide and conquer rule along with a reformed K-Means algorithm which is based on the extracted keywords using cosine similarity is proposed.

Index Terms— Document Clustering, K-Means, Cosine Similarity, Feature Extraction, Weighted K-Means, Reformed K-Means, Topic clustering.

1 INTRODUCTION

This paper is all about clustering text documents comprising multiple topics.

The process of grouping similar entities is called clustering. The similarity between entities is measured through some similarity function that is relevant to the respective domain of entities and depends on the features that decide the similarity or dissimilarity of those entities.

Text mining or document clustering is one challenging domain in itself as the similarity or distance measure is very straight forward. Unlike numerical information, words have different meaning and inference based on the context there is being used and the degree of the emphasis also changes when they are associated with other adjectives.

However this is addressed long ago through cosine similarity which is expressed as a variant of dot product to vectors where each vector representing one document expressed as a directional set of terms associated with their term frequencies.

The challenge left is that one document can comprise of various topics and that makes it hard to be merged with any one clustering that represents one concrete topic. And this is addressed in this paper using an approach called Reformed K-Means.

The Reformed K-Means approach is all about identified the most valued keyword from each passage to be clustered. And this is identified by computing the term document frequencies of each word and then the maximum tfidf associated word is chosen as the keyword. This approach apart being very fast when compared to the existing algorithms also yields a good F-Measure score.

The rest of the paper is organized as: the section 2 speaks about related work, section 3 speaks about the algorithm and scoring methods, section 4 speaks about experimental results followed by the conclusion and references sections.

2 RELATED WORK

2.1 Document Clustering

Grouping documents [1] in such a way that documents belonging to one group is more similar to one another and are far dissimilar from other documents belonging to other groups called clusters.

2.2 The Inspiration

The day by day information analysis is toddled by exponentially increasing documents count. For fast and efficient search, segregation and other applications like opinion mining, behavior mining, it is very important the documents are clustered on the basis of limited time factor and that not compromising in cluster quality which can be measured by the F-Measure.

2.3 Earlier works in Document Clustering

K-Means clustering is a partitioning clustering algorithm which groups given data into K clusters. Algorithms like Novel

algorithm for automatic clustering, Improved partitioning K-means algorithm, Ontology based K-Means algorithm are various variants of the K-Means application for document clustering.

In "A Novel Algorithm for Automatic Document Clustering" [2] , a novel algorithm which generates number of clusters automatically for any unknown text dataset and which clusters the documents appropriately has been developed.

In "A New Partitioning Based Algorithm For Document Clustering" [3] used unsupervised feature selection method to reduce the dimension of document feature space and then proposed a novel partitioning based algorithm which select initial cluster centroids in the process of clustering by the size and density of cluster in the datasets.

"Clustering Technique in Data Mining for Text Documents" [4] describes a system that is designed to identify the semantic relations using the ontology. The ontology is used represent the term and concept relationship. The synonym, meronym and hypernym relationships are represented in the ontology. The concept weights are estimated with reference to the ontology. The concept weight is used for the clustering process. Statistical methods are used in the text clustering and feature selection algorithm.

"Textual Document Clustering using Topic Models" [5] paper introduced a simple extension of naïve LDA clustering method that uses the keyword count matrix and the topic parameters to extract the coordinates of documents for clustering and proposed several simple clustering methods based on the basic topic model.

In each of the above mentioned algorithms, the k-value play a very important role in guiding the algorithms and its results and is a hinder for automatic and fast portioning of text documents and this is the issue we wish to discuss in this paper.

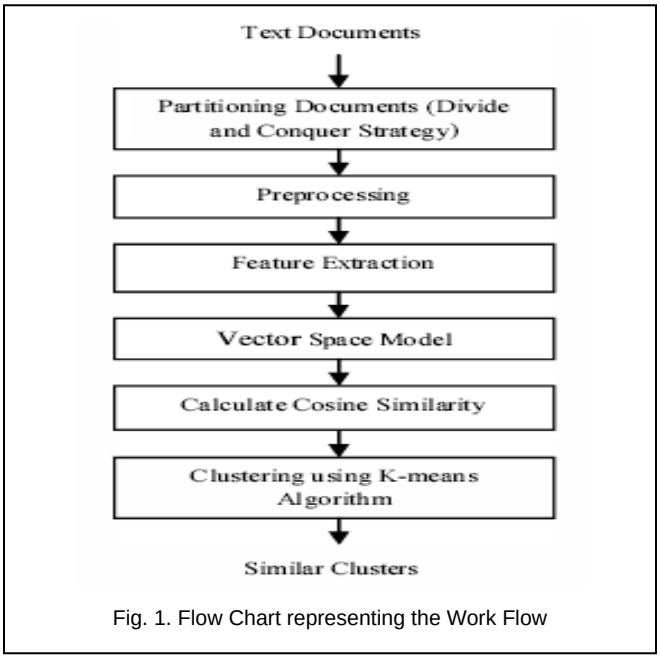

Fig. 1. Flow Chart representing the Work Flow

3 Reformed K-Means For Document Clustering

3.1 Work Flow

The proposed method of document is as follows

1. Input a set of text documents
2. Partition them as groups by topic (Divide and conquer strategy).
3. Pre-process each document by
4. Compute cosine similarity Matrix.
5. Apply K-Means
6. Output resultant clustering with their F-Measure.

3.2 Partitioning By Divide and Conquer

As the previous variant of K-Means was inefficient when the set of documents is large, the set of documents called the corpus is first divided into small sets of documents and then each set is individually portioned.

After the cycle of K-Means is accomplished the resulted clusters from each iteration of partitions is merged to complete the clustering process.

3.3 Pre-Processing of each document

Before applying K-Means on each partition of document, they are preprocessed and expressed as vector space models. The preprocessing is done in five small steps namely:

1. Eliminating specials characters and unwanted symbols - **FILTERING**
2. Splitting sentences and phrases into words called terms – **TOKENIZATION**
3. Eliminating stop words – **STOP WORD REMOVAL**
4. Identifying root words and reducing different forms of such words into their root words and counting their frequencies simultaneously – **STEMMING**.
5. Eliminating word that are under minimum support value - **PRUNING**

3.4 Feature Extraction

As a result pruning terms occurring for not more than 5 times in each document get eliminated and the left out words decide or depict the document more accurate and hence there are called the document features.

3.5 Expressing Document as Vector Space Model

Document Vector Space Model is also known as Term Frequency Inverse Document Frequency Model. A document here is expressed a vector set of elements where each element comprises of the term t and its frequency freq and its term frequency tf and its term weight **w** like a document ***d*** is ***expressed*** as in <(t,freq,tf,w),(t,freq,tf,w),..>.

3.6 Frequency (freq) and Term Frequency (tf)

Number of times a term (word) appears in a document is called its frequency and the term frequency is the normalized frequency of a term with respective to the maximal frequency in that documents.

$$tf(i,j)=\frac{freq(i,j)}{Max\{f(x,j):w\in j\}}$$

Where,
i= term or keyword in document j.
x= Any Term with maximum frequency.

3.7 Inverse Document Frequency (*idf*) and Weight (w)

If in the document corpus we have ***D*** number of documents, and idf is document frequency of a term that is the number of documents containing the term (i) then Inverse Document Frequency ***idf*** and weight ***w*** is

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

$$Wi=tfi*\log(D/dfi)$$

3.8 F-Measure

F-Measure is a metric we employee in this experiment to estimate the quality of each cluster. It is used to compare how similar two clusters are. It is given by,

$$\textbf{F-Measure} = \frac{2\ Precision.Recall}{Precision+Recall}$$

Where,

$$\textbf{Precision} = \frac{|\ relevant\ documents| \cap |retrieved\ documents|}{|retrieved\ documents|}$$

$$\textbf{Recall} = \frac{|relevant\ documents| \cap |retrieved\ documents|}{|relevant\ documents|}$$

3.9 Cosine Similarity

As it is a fact that a document is a collection of passages where each passage is a set of expression through sentence and phrases and each of these have a specific weight age. Considering the weight age of each passage to identify the most possible topic a document belong to is almost critical. Unlike numerical data it is very hard to face the fact that document similarity measure is all about the similar set of words being shared by the given two documents and that the words frequency quantizes the documents topic.

Cosine similarity is expressed as the dot matrix of the vector set expressed documents (say ***A*** .and ***B***)
Where Ai and Bi are components of vector A and B respectively

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

3.10 Reformed K-Means Applied Document Clustering

Putting all together we have the below algorithm.

Input: Dataset set D = {d_1, d2… d_n}
Output: Set of Cluster Numbers C along with document numbers m associated.

1. U= {DI I i € N}
2. Distribute the documents into groups using Dividing and Conquer Merge sort strategy.
   2.1 Apply Divide Strategy on Input Corpus.
   2.2 Apply Divide Strategy till documents are equally placed in groups.
   2.3 Go to step 3.
   2.4 Conquer the Clusters obtained in step 7.
3. Now apply K- Means algorithm on every partition iteratively till we get the same clusters.
4. Calculate the similarity of the documents using cosine similarity measure.
   4.1 Similarity of the document in step 4 is calculated as.
   4.2 for S=Di | i € N Where, D- Documents, N- Number of documents.
   4.3 for i=1 to n Cosine Similarity Matrix

$$CS\ ixj = \begin{pmatrix} 1 & D(1,2) & D(1,3) & D(1,4) \ldots D(1,n) \\ D(2,1) & 1 & D(2,3) & D(2,4) \ldots D(2,n) \\ D(3,1) & D(3,2) & 1 & D(3,4) \ldots D(3,n) \\ \vdots & \vdots & \vdots & \vdots \\ D(n,1) & D(n,2) & D(n,3) & D(n,4) \ldots D(n,n) \end{pmatrix}$$

5. Assign the nearest (similar) document to the new clusters.
6. If the clusters are not matched then go to step 4.
7. If clusters are matched then stop.
8. Conquer the Clusters using Conquer Strategy mentioned in Step 2.4

4 Results And Inferences

4.1 Testing the data

We have developed a java based application that took a 20newgroup dataset as input and applied the proposed reformed K-Means algorithm to cluster that given data set.

Proposed algorithm performs much faster than existing algorithm. Existing algorithm takes more time because it calculates cosine similarity between every documents present in input corpus. Proposed algorithm takes less time because it calculates cosine similarity between the documents present in

every partition simultaneously. Since our proposed algorithm partitions the documents into small set of documents, less number of documents are present in single partition.

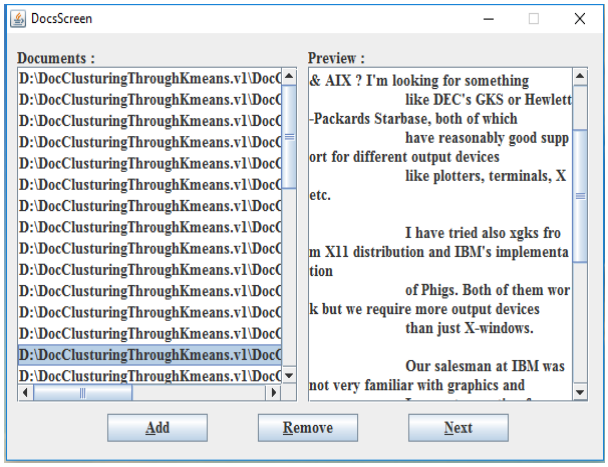

Fig. 3. The Screen Shot Providing Input Corpus Partition

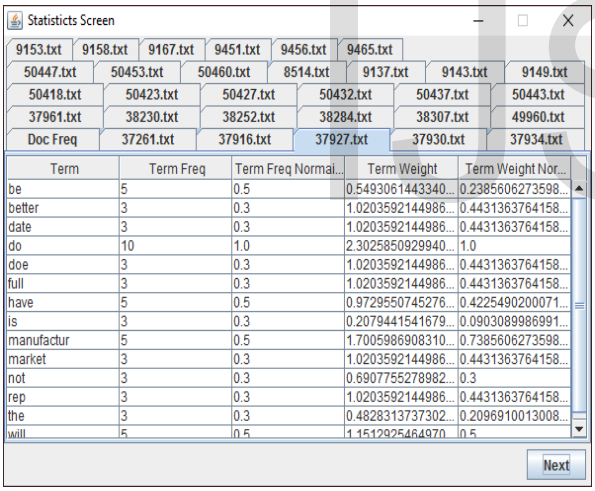

Fig. 4. The Screen Shot showing Frequencies and Term Frequencies and Term Weights of a Document VSM.

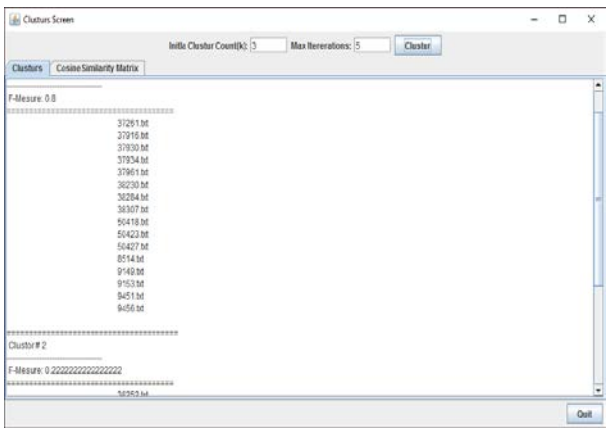

Fig. 5 The Screen Shot showing the final clusters resulted

4.2 Graph depicting F-Measure and Time Complexity

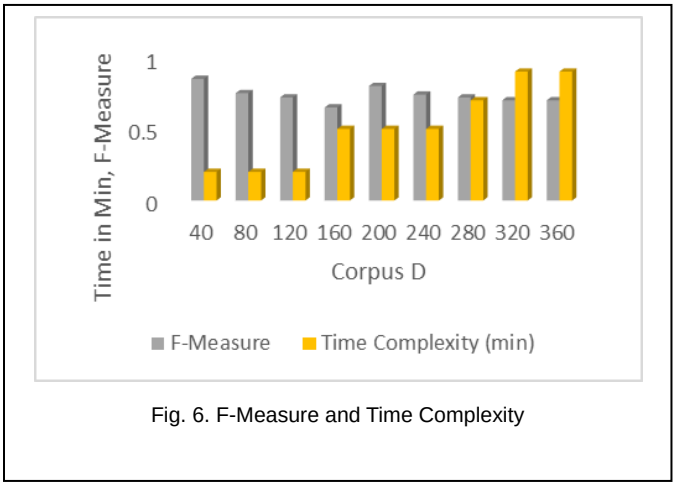

Fig. 6. F-Measure and Time Complexity

5 Conclusion

Clustering documents is a very important process in selecting required documents from a large set of documents. Different variety of K-Means document clustering algorithms lack automation and often perform over clustering. Our proposed reformed K-Means enables automation of clustering and by divide and conquer strategy the K-Means is fed with smaller document groups and the iterative results are then merged to over come over clustering problem. The experimental results have shown that the reformed K-Means algorithm over traditional algorithms have better time complexity and yield very convincing clusters.

6 References

[1] Sargur Shrihari and Graham Leedam , "A survey of computer methods in forensic document examination" Proceedings of the 11 conference of the international Graphonomics society.

[2] Zonghu Wang, Zhijing Liu, Donghui Chen, Kai Tang,"A New Partitioning Based Algorithm For Document Clustering",Eighth International Conference on Fuzzy Systems and Knowledge Discovery,pages 1741 - 1745 IEEE,20 I I.

[3] S.C. Punitha, R. Jayasree andDr. M. Punithavalli, "Partition Document Clustering using Ontology Approach", Multimedia and Expo, 2013 International Conference on Computer Communication and Informatics (ICCCI -2013), Jan. 04 06,pages 1-5, 2013.

[4] J.SatyaPriya,K.Priyadarshini, "Clustering Technique in Data Mining for Textdocuments", IJCSIT, Vol. 3 (1) , 2012.

[5] Xiaoping Sun, "Textual Document Clustering using Topic Models", Knowledge Grid Group Key Lab of Intelligent Information Processing Institute of Computing Technology,

Chinese Academy of Sciences.

[6] Lus Filipe da Cruz Nassif and Eduardo Raul Hruschka, "Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection," IEEE transactions on information forensics and security, Vol. 8, NO. I ,pages 46 - 54 Jan 2013

[7] ROLF OPPLIGER AND RUED! RYTZ, "Digital Evidence: Dream andReality" Swiss Federal Strategy Unit for Information Technology.

[8] Ahmad Mehrbod, Aneesh Zutshi and Antnio Grilo "A Vector Space Model Approach for Searching and Matching Product E-Catalogues" Proceedings of the Eighth International Conference on Management Science and Engineering Management, Advances in Intelligent Systems and Computing, Volume 281 ,pp 833-842,Springer,20 14.

[9] Mushfeq-Us-Saleheen Shameem,Raihana Ferdous, "An efficient K-Means Algorithm integrated with Jaccard Distance Measure for Document Clustering", pp 1-6,IEEE,2009.

[10] V. Mary Amala Bai,Dr. D. Manimegalai , lCCCCT-IO, "An Analysis of Document Clustering Algorithms",pp402-406, IEEE, 2010.

[11] Sonia Bui,Michelle Enyeart, Jenghuei Luong, "Issues in Computer Forensics" COEN 150, November 2003.

[12] Xiaoping Qingl, Shijue Zheng ,"A new method for initialising the K-Means clustering algorithm" Second International Symposium on Knowledge Acquisition and Modeling,pp 41-44,IEEE,2009.

[13] Ranjana Agrawal ,Madhura Phatak, "A Novel Algorithm for Automatic Document Clustering," 3rd IEEE International Advance Computing Conference (IACC) ,pages 877 - 882,IEEE, 2013.

[14] M.K.V.Anvesh and Dr.B.Prajna "Potential based similarity metrics for implementing hierarchical clustering", 103-volume-4-issue ,IJECS ,2015

[15] Tapas Kanungo,David M. Mount,Nathan S. Netanyahu,Christine D. Piatko, Ruth Silverman, and Angela Y. Wu "An Efficient k-Means Clustering Algorithm: Analysis and Implementation" IEEE transactions on pattern analysis and machine intelligence, vol. 24, NO. 7,pp 881-892 JULY 2002.

[16] K.Naga Neerja, B.Prajna , "An effective Research Paper Recommender System based on Subspace Clustering", International Journal Of Engineering And Computer Science, Page No. 13306-13310 Volume 4, Issue 7, July 2015 .

[17] Elizabeth Len, Jonatan Gmez and Olfa Nasraoui,"A Genetic Niching Algorithm with Self-adaptating Operator Rates for Document Clustering", Eighth Latin American Web Congress, 2012.

[18] A. E. ELdesoky, M. Saleh, N.A. Sakr, "Novel Similarity Measure for Document Clustering Based on Topic Phrases",pages 92 96,1 EEE,2009.

[19] Sargur Shrihari AND Graham Leedam , "Study of Clustering Algorithm based on Model data" Proceedings of the II conference of the international Graphonomics society, pages 3961 - 3964 November 2007

[20] Nadempalli Sneha, B.Prajna , Sharmila Sujatha , "Application for Retriving Details of Users - Topic Based Approach", pages 509-513 ,volume 6, Issue 8,IJCSET,2015

[21] B.Prajna , Shashi M , "Document Clustering Technique based on Noun Hypernyms", IJECT Vol. 2, SP-1, Dec. 2011.