International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307

Fuzzy Document Clustering based on Frequent Features and Feature Length

U. S. Patki*[1], Dr. S. B. Kishor[2], Dr. P.G. Khot[3]
[*1]Department of Computer Science, Science College Nanded, Maharashtra, India
[2]Department of Computer Scinece, S.P. College, Chandrapur, Maharashtra, India
[3]Ex-Professor, Department of Statistics, RSTM Nagpur University, Nagpur, Maharashtra, India

ABSTRACT

Document Clustering is a method of grouping similar documents into one cluster. Fuzzy document clustering is a soft computing technique used for clustering the similar documents. It permits overlapping i.e. it permits single document to belong to multiple clusters. Feature selection and feature extraction is the most important phase during clustering process. In the Literature different feature reduction methods are proposed. In this research paper we have proposed a feature reduction method based on feature frequency and feature length. In this method, we have chosen the features based on no. of occurrence in a set of N documents. We have also taken into account feature length. Finally we have applied fuzzy C-Means clustering algorithm for clustering the N documents into K-Clusters.

Keywords: Document Clustering, Soft Computing, Features selection, features reduction, Fuzzy C-Means.

I. INTRODUCTION

Today information is rapidly growing on Internet. This information is available in the form of Text, Images, Graphs, Charts, and Tables etc. To search the exact information has becoming a challenging task. If the documents containing textual information are well grouped according to their contents, searching becomes easier task.

In the Literature, different methods for document clustering are proposed. The purpose of each method is to group the documents into clusters based on their contents. These methods includes Hierarchical clustering, partitional clustering like K-Means, C-Means, Soft Clustering like Fuzzy C-Means etc.

What is Document Clustering?

[1] We have given a set of N documents. Our objective is to extract the contents of each documents, find the most important words which we refer it as features and group these documents into K-clusters such that:

i) Intra-cluster distance is minimum
ii) Inter-cluster distance is maximum

The Paper organized as follows. In the introduction part we have introduced the research topic, we have the described the steps for document clustering. It is followed by Feature selection methodology we have proposed. We have finally presented some experimental results with a conclusion.

II. STEPS IS DOCUMENT CLUSTERING

[2]The Different steps involved in document clustering are as shown in the figure bellow. It consists of five important steps viz. pre-processing, feature extraction, feature selection, clustering method and analysis of results.

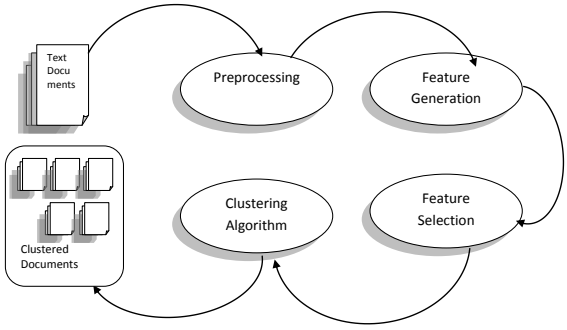

Fig1. Steps in Document Clustering

1. Pre-processing: Pre-processing is the step of cleaning the text documents. It involves punctuation symbols and stop-words removal, Stemming process and term weight assigning.
2. Feature Generation: In this step features are generated. It includes generation of vector space matrix using features.
3. Feature selection: It involves dimension reduction of vector space matrix. In this step features which do not have too much importance are removed.
4. Clustering Technique: This is the final step which actually clusters the documents into specified number of groups.
5. This involves verification and analysis of clustering results.

III. FEATURE SELECTION METHOD

This step is carried out after pre-processing and feature generation. In step is also known as feature or dimension reduction. This step is applied on vector space model (VSM) which is generated in the first two steps of clustering. The vector space model represents M features that occur in N documents. Hence the size of vector space model will be N x M
As the number of documents will be increased, the number of features will be also increased. It will rapidly increase the size of VSM. In the literature, different algorithms are proposed for feature reduction or dimension reduction.

[3] An author(A. Sudha Ramkumar, Dr. B Poorna, November 2016) suggested different dimension reduction techniques based on Document frequency, Information Gain, Mutual information, Chi Square Statistics etc. Author says that "Information gain is an effective Feature selection method and widely used method in Text data mining."

[4] An Author (Ammar Ismael Kadhim, Yu-N Cheah and Nurul Hashimah Ahamed, 2014) proposed another dimension reduction technique using Single Value Decomposition (SVD). Author proposed decomposition of vector space matrix into a correlated set of singular values and two orthogonal bases of Singular vectors as bellow:

$$H=USV^T$$

Where U and V is unitary matrix of dimension n X n and m X m. S is an n X m diagonal matrix and T defines transposition.

[5] According to author (MS K. Mugunthadevi. MRS S.C. Punitha, Dr. M. Punithavalli, 2011) "Feature selection algorithms typically fall into two categories: feature ranking and subset selection. Feature ranking ranks the features by a metric and eliminates all features that do not achieve an adequate score. Subset selection searches the set of possible features for the optimal subset".

IV. FREQUENT FEATURES AND FEATURE LENGTH

In our proposed method we have considered only those features that occur in at least k documents. If a feature occurs in at least k documents then and then the feature will be included in vector space matrix. The second criterion that we have imposed is the feature length. We have assumed only those features whose length is at least three characters. i.e.

$$\|W_{ij}\| \geq 3$$

where W_{ij} is j^{th} feature of i^{th} document.

Using the above two criterions we have performed our experiment on the standard database 20Newsgroup database.

V. EXPERIMENTAL RESULTS

In our experiment we have considered three different classes of datasets from 20Newsgroup database. First class contains Class-1 contains 50 documents with 2 Clusters, Class-2 contains 60 documents with 2 Clusters and Class-3 contains 120 documents with 3 Clusters.

We have used Matlab 13(b) tool for our experiment. The table given bellow shows the details of our experiment.

Table 1. Summery of experiment

Class No.	1	2	3
Total Doc	50	60	120
Feature Length	3	3	3
Features in Min Docs	5	2	5
Total Words	39286	33568	88770
Total Unique words	2987	3535	5499
Total Features Selected	1668	1817	3278

The table 1 shows that using the frequent features and feature length approximately 50% features are reduced. The fig bellow shows the experimental results using fuzzy document clustering algorithm on our experimental database as shown in table 1.

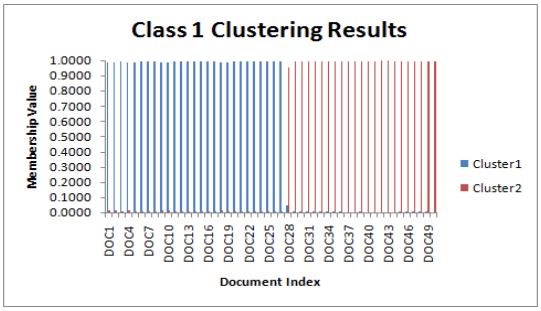

Figure 2. Results for Class-1

Tabel 1

	Actual Documents	Result Shown	% Result
Cluster 1	25	29	86.21%
Cluster 2	25	21	84.00%

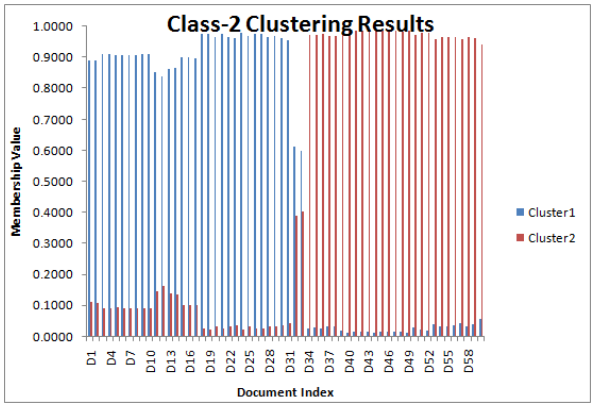

Figure 2. Results for Class-2

Table 3

	Actual Docs	Result Shown	Overlapping
Cluster1	30	31	02
Cluster2	30	17	02

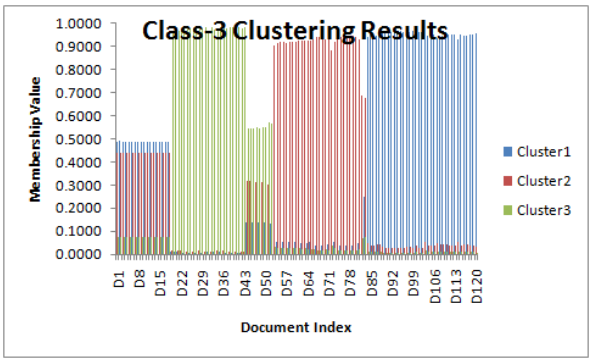

Figure 3. Results for Class-3

Table 4

	Actual Docs	Result Shown	Overlapping		
			C1+C2	C2+C3	C1+C3
Cluster	40	39	18	09	Nil
Cluster	40	36	18	09	Nil
Cluster	40	27	18	09	Nil

In the class-1 & 3 we have used only those feature words that occur in at least *five* documents. It gives better results as compare to class 2 where we have choose feature words that occurs in at least *two* documents. In all three cases, we have assumed word length as three. This criterion has reduced approximately 50 percent of feature words without affecting the result.

VI. CONCLUSION

Feature word selection is a very important step in document clustering. The features that are most frequently occurs in the vector space matrix can be used in the clustering process. These features must occur in at least K documents. The value of K depends on the total number of documents to be cluster.

We have also the feature length as a feature selection criterion. It is very common thing that the words whose length in less than three characters does not have too many important in our clustering process. Hence we have ignored such words. It has reduced the feature matrix dimension.

This feature selection concept can be further extended by assuming synonyms in the feature words. It will further reduce the dimensions of feature matrix.

VII. REFERENCES

[1]. Sumit Goswami and Mayank Singh Shishodia, "A Fuzzy Based Approach To Text Mining And Document Clustering"2013

[2]. Sowmya P, Supreetha R,Ushadevi A, "Survey On Algorithms Used for Text Document Clustering", IJAEC Special Issue September 2016

[3]. A. Sudha Ramkumar, Dr. B Poorna. (November 2016). Text Document Clustering Using Dimension Reduction Technique. International Journal of Applied Engineering Research , 4770-4773.

[4]. Ammar Ismael Kadhim, Yu-N Cheah and Nurul Hashimah Ahamed. (2014). Text Document Preprocessing and Dimension Reduction Techniques for Text Document Clustering. IEEE Computer Society (pp. 69-73). IEEE.

[5]. MS K. Mugunthadevi. MRS S.C. Punitha, Dr. M. Punithavalli. (2011). Survey On Feature Selection in Document Clustering. International Journal on Computer Science and Engineering (IJCSE) , 12401-1244.

[6]. Anna Huang, ," Similarity Measures for Text Document Clustering", NZCSRSC 2008, April 2008, Christchurch, New Zealand

[7]. A Text Book " Text Mining and Application Programming" Manu Konchady ,Ed. 3 Indian Edition