A Fuzzy based Approach for text document clustering

Mausumi Goswami
Assam University, Silchar

Bipul Syam Purkayastha
Asam University, Silchar

Abstract— **As the amount of information is increasing exponentially, so is the number of users day by day. There is tremendous increase in the volume of text documents available on the World Wide Web such as in news sites, intranets, extranets, and digital libraries. Clustering is done to group similar items based on a similarity criterion. Soft computing approach may be applied to perform clustering. An indistinct way of clustering may be combined with kmeans. This combined algorithm works similar to kmeans but produces indistinct membership. The Fuzzy c-Means algorithm is a blend of k-means and fuzzy logic techniques where each document may belong to more than one group. For every document the extent of membership or degree of membership is calculated for all the clusters. In this paper a fuzzy based approach for text document processing is discussed.**

***Index terms*— Stop words, Boolean, fuzzy c-means, cosine similarity.**

I INTRODUCTION

Web crawling requires to categorize bulk documents with respect to certain specific criteria. In our day to day life information retrieval is done in various forms- while searching web, searching email or searching a collection of stored documents. Text document clustering plays an important role in information retrieval. Text mining is about knowledge discovery from large collections of unstructured text. It’s not the same as data mining. Data mining is more about discovering unknown patterns in structured data stored in databases.

Unstructured nature of text causes many problems in text mining although there are lot of existing similar kind of techniques available in data mining. Apart from this a major constraint is the human speaking language or natural language. When facts and structured information are extracted from unstructured text the process is called information extraction (IE). The extracted facts are analyzed. Retrieving documents from large text collections i.e the worlds wide web in response for a given query or specific keywords or index words. Analysis is performed on the retrieved documents. Structured information is returned by IE whereas IR returns documents containing the relevant information.

Different stages of text mining are document selection and filtering (IR techniques) , document pre-processing (Natural Language Processing techniques), document processing (NLP / ML / statistical techniques). Text document clustering is grouping of similar documents into meaningful clusters. Text document clustering is grouping of similar documents into meaningful clusters.

A Document Clustering Problem can be defined as below:

Given (i) a set of documents D = d1,d2,d3,…… dN,

(ii) a desired number of clusters k, and

(iii) an objective function f that evaluates the quality of clustering, we want to compute an assignment of documents d1,d2 etc into clusters k1,k2 etc that minimizes (or, in some cases, maximizes) the objective function. Mostly, the mapping is surjective which means none of the K clusters is empty. We may define the objective function in terms of a similarity measure or distance function. Different sections described in this paper are the following: Section II briefs about pre-processing of documents and existing representation models, Section III about similarity measures , Section IV about clustering and results, section V about conclusion.

II PRE-PROCESSING THE DOCUMENTS

When a document is fed into an information retrieval system few steps of pre-processing is required on the document. The steps required to be followed can be divided into five levels:

1. Lexical analysis
2. Stop word elimination
3. Stemming
4. Index-term selection
5. Construction of thesauri

Stemming improves s*torage and search efficiency* since less terms are stored. The processing of the documents is a very important process, which keeps only the relevant data and the unnecessary data are removed. It makes the search easier and it is also used in document classification.

Removal of duplicate or redundant data and special characters are done in document preprocessing. It also involves removal of stop words and stemming.

Steps involved in this process are mentioned below:

1. Read the input documents from files
2. Read the set of stop words from file
3. Parse the input document into words of size length more than 1
4. Compare each word with defined set of stopwords
5. Filter the stop words and retain other words

III REPRESENTATION MODELS

There are different ways to represent a collection of words. The most basic one is called Boolean Model. The other models are variation of Tf-IDF models. Term frequency–inverse document frequency is a score given to measure the weight of the word in the collection of documents. In this paper TF-IDF model is used .

IV SIMILARITY MEASURES

Coherent patterns are grouped together to form a cluster by text document clustering. Documents that are different get separated apart into different clusters. The similarity among documents varies depending on the context. If we consider clustering research papers i.e. research papers are treated as input documents, two documents are considered as similar if they share similar thematic topics. Again when clustering web sites more emphasis is put on the type of information that is presented in the page. This kind of clustering can benefit further analysis and utilization of the dataset such as information retrieval and information extraction. This is done by grouping similar types of information sources together. A suitable and precise definition of the closeness between a pair of documents is required to perform accurate clustering. The closeness or similarity between documents may be measured in terms of similarity or distance function. Understanding and realizing the effectiveness of different measures is of great importance to choose the most suitable similarity measure or distance function. A variety of similarity or distance measures have been proposed and widely applied, such as Euclidean Distance, Cosine Similarity, Pearson Correlation Coefficient, Jaccard coefficient, and averaged Kullback Leibler Divergence. Similarity measures which have been frequently used for document clustering are discussed below.In this paper we considered cosine similarity.

V CLUSTERING AND RESULTS

One of the most popular soft clustering techniques is Fuzzy c-means clustering. It is a combination of K-means and Fuzzy Logic Technique. The fuzzy c-means (FCM) algorithm is a clustering algorithm developed by Dunn, and later on improved by Bezdek is useful in the following situations:

i) The number of clusters are known
ii) Calculates the degree of membership of a document to all the clusters.
iii) Does not focus on absolute membership
iv) Fast because the number of iterations required

to achieve a specific clustering depends on the required accuracy and membership is not hard.

During FCM a membership matrix is produced which contains degree of membership of a document to all the clusters. A N X K membership matrix called U is produced with N documents and K clusters. Each entry of this matrix $u_{i,j}$ lies between 0 and 1. The sum of all the elements in a specific row is 1.

There are two input parameters for the fuzzy clustering module. Input parameters are the following:

1. Input data matrix (D): This is a matrix where every row represents a sample data point.
2. N: the number of predefined clusters.

The output parameters obtained from this module are the following:

1. Final_cluster_centers (center): This is a matrix where each row gives the cluster center coordinates. i.e if there are 5 clusters then there will be five rows , each one for every cluster. Each row will be of 200 points showing the center of the corresponding cluster.
2. Membership_matrix : This is the final membership matrix or fuzzy partition matrix which shows the membership value for every document in each cluster.
3. objective_fcn: values of the objective function during iterations

The fuzzy c means algorithm is applied to a given data set of size 100 x 200 containing 100 documents and 200 terms in each documents. The number of clusters are varied and results included.

We may consider one more optional argument variable ‘op’ as input parameter to specify clustering parameters which can introduce a stopping criteria, or set the iteration information display. The default value for maximum number of iteration is 100. The default value for minimum amount of improvement is 1e-5. In this method the stopping criterion for the clustering process is decided based on maximum number of iteration or accuracy level. By accuracy level it is meant that when the objective function improvement between two consecutive iterations is less than the minimum amount of improvement specified.

The algorithm is implemented with 100 documents . In the following diagram the output obtained in a machine with intel pentium quad core processor, windows 8.1 using MATLAB R2012a is shown below. The experiment is conducted with varied number of clusters. For five clusters and 100 documents the following result is obtained:

Iteration count = 1, obj. fcn = 446.937602
Iteration count = 2, obj. fcn = 330.561669
Iteration count = 3, obj. fcn = 330.351286
Iteration count = 4, obj. fcn = 330.350124
Iteration count = 5, obj. fcn = 330.350115

Table: for 5 cluster and 100 documents the values obtained for objective function.

The same algorithm is implemented with 4 clusters and 2 respectively. The snapshot of the results is tabulated below:

Iteration count = 1, obj. fcn = 545.705864
Iteration count = 2, obj. fcn = 411.448573
Iteration count = 3, obj. fcn = 411.168287
Iteration count = 4, obj. fcn = 411.166624
Iteration count = 5, obj. fcn = 411.166611
Iteration count = 6, obj. fcn = 411.166610

Table: for 4 cluster and 100 documents the values obtained for objective function

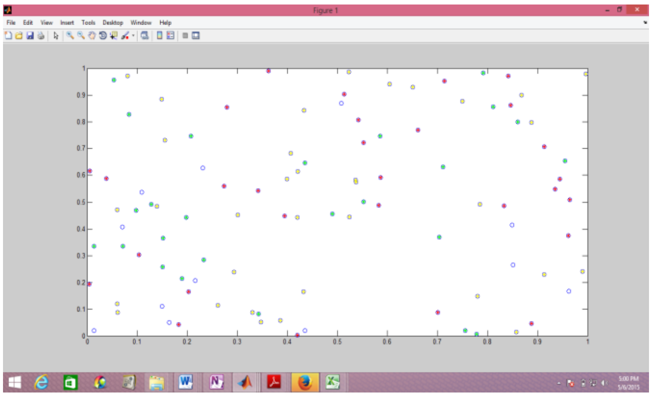
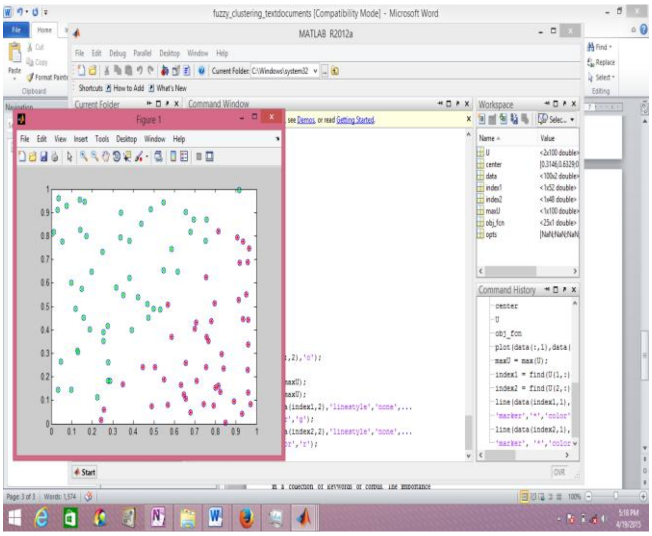


Figure: 4 clusters , 100 documents

In the above figure four different clusters are given by four different colors red, green, yellow and red.



Name	Value
U	<2x100 double>
center	[0.3146,0.6329;0
data	<100x2 double>
index1	<1x52 double>
index2	<1x48 double>
maxU	<1x100 double>
obj_fcn	<25x1 double>
opts	[NaN;NaN;NaN

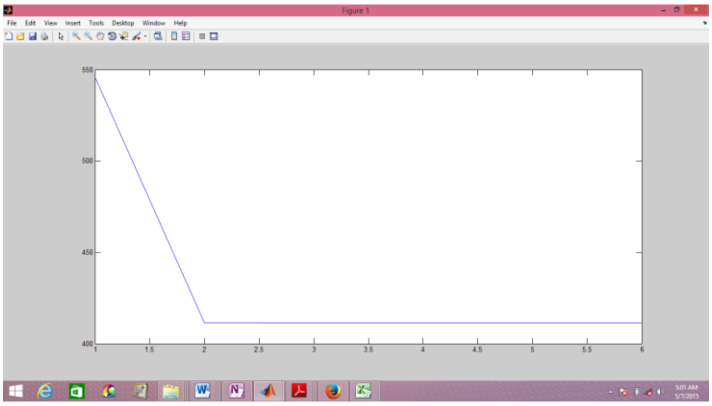
```
:,2),'o');

maxU);
maxU);
a(index1,2),'linestyle','none',...
r','g');
a(index2,2),'linestyle','none',...
r','r');
```

```
center
U
obj_fcn
plot(data(:,1),data(
maxU = max(U);
index1 = find(U(1,:)
index2 = find(U(2,:)
line(data(index1,1),
'marker','*','color'
line(data(index2,1),
'marker', '*','color
```

Figure: Fuzzy c-means clustering results for two clusters with 100 documents . In this case red and green colors are used to indicate two different groups.

In the following figure the change of objective function values is demonstrated.



V Conclusion

In this paper a fuzzy based approach for text document clustering is implemented. The number of text documents is varied and number of clusters is also varied. The objective function is studied and graph is drawn. In our future work a comparative study of fuzzy based c means and kmeans will be implemented. Also the impact of considering various similarity measures will be compared.

References

[1] Mausumi Goswami, Gowtham, Balachandran, B. Purkayastha, “An Approach for Document Pre-processing and K Means Algorithm Implementation”, IEEE conference held at Kochi in August,2014

[2] Mausumi Goswami, B.Purkayatha,”Term frequency and inverse document frequency based preprocessing” , international conference IAETSD: ICDER – 2015

[3] Stuti Karol, Veenu Mangat, “Evaluation of text document clustering approach based on particle swarm optimization” Springer ,September,2012

[4] Bezdek E., Full, “FCM: the fuzzy c-means clustering algorithm”, Comput. Geosci., 10, 191-203, 1984

[5] Anna Huang, “Similarity Measures for Text Document Clustering”,*NZCSRSC 2008*, April 2008, Christchurch, New Zealand

[6] https://www.mediafire.com/folder/mfk5wjfbj8lwp/Data_sets