I.J. Mathematical Sciences and Computing, **2015, 4, 34-43**
Published Online November 2015 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijmsc.2015.04.04

***Available online at** http://www.mecs-press.net/ijmsc*

A Fuzzy Approach for Text Mining

Deepa B. Patil[a], Yashwant V. Dongre[b]

[a]*Vishwakarma Institute of Information Technology,3/4 Kondhwa (Bk), Pune-411048, India*
[b]*Vishwakarma Institute of Information Technology,3/4 Kondhwa (Bk), Pune-411048, India*

Abstract

Document clustering is an integral and important part of text mining. There are two types of clustering, namely, hard clustering and soft clustering. In case of hard clustering, data item belongs to only one cluster whereas in soft clustering, data point may fall into more than one cluster. Thus, soft clustering leads to fuzzy clustering wherein each data point is associated with a membership function that expresses the degree to which individual data points belong to the cluster. Accuracy is desired in information retrieval, which can be achieved by fuzzy clustering. In the work presented here, a fuzzy approach for text classification is used to classify the documents into appropriate clusters using Fuzzy C Means (FCM) clustering algorithm. Enron email dataset is used for experimental purpose. Using FCM clustering algorithm, emails are classified into different clusters. The results obtained are compared with the output produced by k means clustering algorithm. The comparative study showed that the fuzzy clusters are more appropriate than hard clusters.

Index Terms: Fuzzy clustering, fuzzy c means clustering algorithm, text mining

1. Introduction

Document clustering or data clustering divides data items into groups so that items in the same group are most similar, at the same time; they are most dissimilar with the data item in other clusters. Depending on the nature of the data and the purpose of clustering, different measures of similarity can be used to place items into different clusters or groups. The main objective of similarity measure is to control cluster formation. Some methods find similarity between two objects by distance between them. Such distances can be defined using Euclidean distance, Cosine similarity, dice coefficient, extended Jaccard coefficient etc.

Data can be partitioned into hard clusters or soft clusters. In hard clustering, data items are divided into separate clusters, where each data element belongs to only one cluster. In soft clustering or fuzzy clustering, data items may belong to more than one cluster, with different degree of membership. For example, if we want

* Corresponding author. Tel.:+91-9822893671
E-mail address: deepa_p100@yahoo.co.in

to partition height of human being as tall, medium and short, we can consider a person as a tall if his/her height is above 5'6", person as medium in height if his/her height falls between 4'6" to 6" and person is short if his/her height is below 5'. A person whose height is 5'9" falls in both groups, in medium as well as in tall. The height 5'9" is closer to the value 6" so the person is classified as tall. Such type of classification is more appropriate and accurate using fuzzy clustering than hard clustering wherein there is a chance of falling data item in wrong cluster. My work described here, focuses on comparative study of hard clustering vs soft clustering using k means clustering algorithm and Fuzzy C Means (FCM) clustering algorithm using five similarity/distance measures namely, Euclidean distance, Cosine similarity, dice coefficient, extended Jaccard coefficient and Similarity Measure for Text Processing (SMTP).

For experimental purpose I used Enron email data set [1] which is which is available free on World Wide Web. I also created a small email data set for comparative analysis. Fuzzy classification can be rule based or keyword based. I focused my work on classifying emails based on keywords. Many a times there is a situation wherein a person is unable to locate a piece of information he/she knows is out but can't find it. This can be extremely frustrating, particularly if you know you've seen it before. Corporate emails, most of the times contain some typical keywords. For example, employees of research and development department of pharmaceutical company may receive emails with some particular keywords describing name of diseases, drugs , contents in drugs , composition, name of another pharmaceutical company, symptoms of disease, side effect of particular drug etc. Finding a particular email or emails containing required name of diseases doesn't require fuzzy rules as fuzzy rules are not applicable in such situations. So my main focus was on keyword based fuzzy classification.

2. Literature Survey

A lot of similarity measures are in existence to calculate similarity between given two documents. Euclidean distance [2] is one of the popular similarity measures. Cosine similarity [3] is a measure which takes cosine of the angle between two given document vectors. The Jaccard coefficient [4] is a statistic used for comparing the similarity of two document sets. It is defined as size of intersection divided by size of union on sample data sets. An information-theoretic measure for document similarity called IT_Sim [5], [6]. It is a phrase-based measure which computes the similarity based on Suffix Tree Document Model. Pairwise-adaptive similarity [7] is a measure which selects a number of features dynamically, out of document d1 and document d2. In [4], [8] Hamming distance is used, hamming distance between two document vectors is number of positions where the corresponding symbols differ. In [9] a non-symmetric similarity measure called Kullback-Leibler divergence is described. It is difference between probability distributions associated with two vectors. In [10] an advanced similarity measure is proposed known as Similarity Measure for Text Processing (SMTP) which gives more value to presence or absence of features (words) than frequency of features (words).

Lotfi A. Zadeh is initiator of fuzzy logic. He introduced fuzzy sets [11] in 1965. Fuzzy sets are based on fuzzy logic. In a keynote speech, Zadeh himself have said that fuzzy logic is not fuzzy, but is a precise logic of imprecision. In fuzzy logic, a variable may have any real value between 0 and 1 unlike in boolean logic where value of variable is either 0 or 1. Fuzzy logic can have linguistic variables, for example age, which may take non-numeric values such as young, middle-aged or old. Amongst early applications of fuzzy logic, notable ones are use of fuzzy logic in high speed train to improve precision of ride and economy, in handwriting recognition and to improve fuel consumption in automobiles. Fuzzy logic has many applications in the field of engineering as well as non-engineering fields. For example, fuzzy logic can be applied in the fields of artificial intelligence, image processing and control theory, medical diagnosis systems, stock trading applications to name a few. Fuzzy logic used in washing machine can be used to control washing process such as intake of water, temperature of water such as hot, cold, lukewarm , cloth wash time, spin speed and rinse performance. Thus, use of fuzzy logic in washing machine helps to increase its lifespan. Zadeh explored more about fuzzy logic [12]. In [13] Zahed et.al contributed more about fuzzy sets and fuzzy logic. Kosko [14] has important contribution in development of fuzzy logic in the field of artificial intelligence.

There are many fuzzy clustering algorithms which are proposed by various researchers, namely fuzzy C-means, fuzzy K-nearest neighbor, fuzzy ISODATA, algorithm, potential-based clustering, and many others [15]. Fuzzy C-means (FCM) clustering algorithm is one of the most popular and widely used fuzzy clustering algorithm. FCM was originally proposed by Dunn [16], later modified by Bezdek [17]. FCM determines, and iteratively updates the membership values of each data point in each of the clusters. So, a data point is member of all clusters with varying degree of membership values. The logic of FCM is extensively used in varied fields of research [18, 19, 20, 21]. There are several variants of FCM algorithm. Sikka et al. [22] developed a modified FCM known as MFCM to estimate the tissue and tumor areas in a brain MRI scan. Krinidis and Chatzis [23] proposed a Fuzzy Local Information C-Means (FLICM) algorithm. A modified FCM algorithm was developed by Belhassen and Zaidi [24] to overcome the problems faced by conventional FCM algorithm .

3. Proposed System

Following figure describes proposed system in detail.

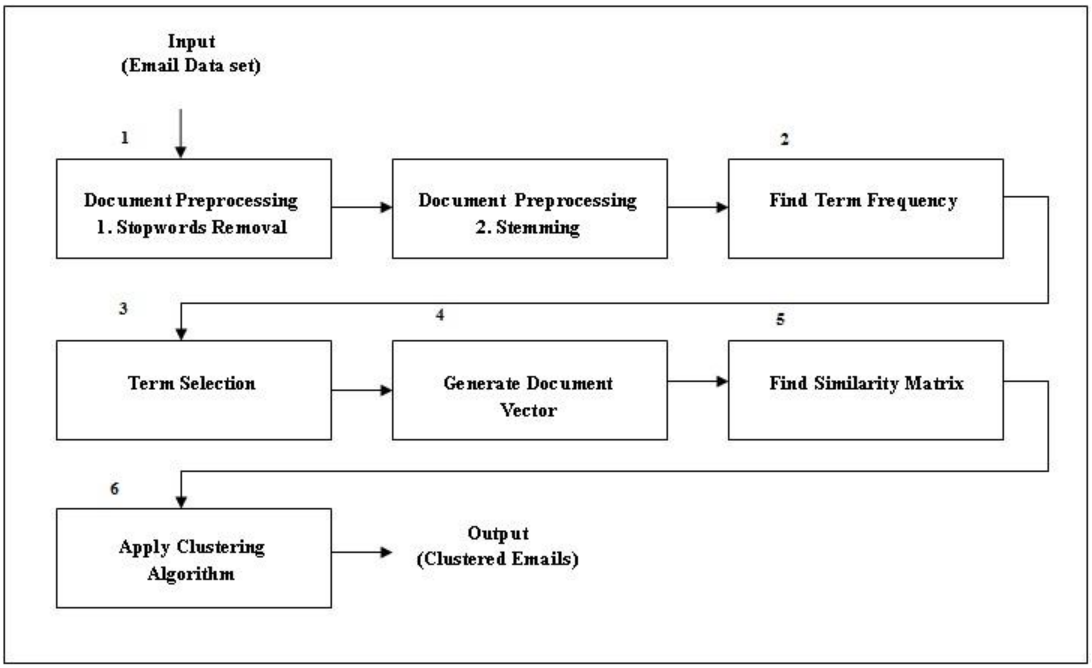

Fig. 1. System Architecture

Following steps will be carried out to achieve final results. In the above figure, the steps are numbered as 1 to 6.

- Document Preprocessing- Document preprocessing can be carried out in two steps namely- Stopwords removal and Stemming
  - Stopwords Removal – Words which occur very frequently are not useful for the purpose of information retrieval. It is observed that a word which appears in almost 80% of documents in the document set, are useless for information retrieval. Such word is referred to as stopword. Example of such words are – a, an, the, are, on, to, about, above, up to, onto, etc. In stopwords removal process, stopwords are removed from the document collection.

- Stemming – Stem is the fraction of word which is left after removal of its affixes. So, stemming is the process of removing plurals, gerund forms and past tenses from a word. An example of stem is the word 'calculate' which is a stem for variants such as calculated, calculating, calculation and calculations.

- Find term frequency- After pre-processing document collection, each word that remains in the document collection is called as 'term'. Frequency of each term in each document in the collection is calculated. For example if an email contains following text-
  "First year engineering admissions (FE) are commencing from 2nd July; so, you are requested to update college website on priority basis. Please update the Placement Report on our College website. Pls do the needful as early as possible as FE admissions are commencing from 2 nd July and people will surf our website more frequently. "
  After stopwords removal and stemming, the document will have following terms-
  first year engineering admission (FE) commenc July request update college website priority update placement report college website needful possible FE admission commenc July people surf website frequent
  The frequency of each term in the document is calculated. In the above example, frequency of word 'website' is 3
- Term selection- All the terms in the collection are not useful for information retrieval. For example terms with low frequency count may not be considered, for example in above example terms such as 'FE' and 'needful' with low frequency count are not useful for information retrieval, so such terms are omitted. In term selection step, terms with high frequency count are retained for information retrieval.
- Generate document vector- In this step, document vector for each document is generated. Consider following two emails –
  Email 1:- First year engineering admissions (FE) are commencing from 2nd July; so, you are requested to update college website on priority basis. Please update the Placement Report on our College website. Pls do the needful as early as possible as FE admissions are commencing from 2 nd July and people will surf our website more frequently.
  Email 2:- Kindly update the designation "Asst. Professor" on related pages on college website. Also add email id in the description deepa.patil@viit.ac.in and wherever needed on our college website. Pls update website as early as possible as DTE visit is scheduled on Friday 27th this month.
  For above two emails, following terms are identified which are displayed below in alphabetical order-
  college, designation, engineering, email, placement, website
  For the above terms, document vector for each email is-
  document vector for email 1:- <2,0,1,0,1,3>
  document vector for email 2:-<2,1,0,1,0,3>
- Find similarity matrix- Calculate similarity matrix for all documents in the document collection using five distance measures namely, Euclidean distance, Cosine similarity, dice coefficient, extended Jaccard coefficient and Similarity Measure for Text Processing(SMTP) with the help of document vector generated for each document in the above step. Separate similarity matrix is generated for each distance measure.
- Apply clustering algorithm-Apply clustering algorithms k means and FCM on each of the similarity matrix generated in the above step. The final outcome is clustered emails. The output produced by two clustering algorithms using above mentioned distance measures are then compared.

4. Algorithmic Details

4.1. K Means Clustering Algorithm

Following are the steps for k means clustering algorithm
Step 1: Arbitrarily choose k objects from data set as initial cluster centers(centroids)
Step 2: Repeat
Step 2.1: Determine the distance using distance measure, between each object and each one of the centroid
Step 2.2: Assign the object to the cluster to which the object is most similar
Step 2.3: Update the cluster centers
Step 3:Until no change

4.2. Fuzzy C Means Clustering Algorithm

The aim of fuzzy c means clustering algorithm is to minimize objective function –

$$J(U,V|Z) = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ik})^m \ \|Z_k - V_i\|_A^2 \quad (1)$$

Following are the steps for FCM clustering algorithm-
Step 1: Initialize the matrix $U=[u_{ij}]$ matrix, $U^{(0)}$
Step 2: At k-step: calculate the centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$ using following equation -

$$C_j = \frac{\sum_{i=1}^{N} u_{ij}^m . X_i}{\sum_{i=1}^{N} u_{ij}^m} \quad (2)$$

Step 3: Update $U^{(k)}$, $U^{(k+1)}$ using following equation -

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|X_i - C_j\|}{\|X_i - C_k\|} \right)^{\frac{2}{m-1}}} \quad (3)$$

Step 4 : If $\| U^{(k+1)} - U^{(k)}\| < \varepsilon$ then STOP; otherwise return to step 2.
In the above algorithm –

- u_{ij} is the degree of membership of x_i in the cluster j
- c_j is the center of the cluster
- This loop iteration will stop when

$$max_{ij} \left\{ \left| u_{ij}^{(k+1)} - u_{ij}^{(k)} \right| \right\} < \varepsilon \quad (4)$$

- **m** is fuzziness coefficient – **m** determines how much clusters can overlap each other. **m** lies between 1 and ∞. Higher the value of **m** , more data points will lie in fuzzy band. Usually initially **m=2** is chosen
- **ε** is termination tolerance – the algorithm stops when $\| U^{(k+1)} - U^{(k)}\| < \varepsilon$. Usually choice of **ε** is .001

5. Result Analysis

5.1. Dataset Used

For experimental purpose, Enron email data set [1] is used which can be downloaded free. The data set is cleaned and made available on the World Wide Web for research purpose. A small email data set is also

created for comparative analysis.

5.2. Analysis

The result analysis is done on the basis of similarity measures and clustering algorithms used. For experimental purpose, four clusters are considered. First, K means clustering algorithm is applied on the similarity matrix generated, for five distance measures namely, Euclidean distance, Cosine similarity, dice coefficient, extended Jaccard coefficient and Similarity Measure for Text Processing (SMTP). Experts' results are generated, that is, emails for each cluster for each similarity measure are identified and are compared with system generated email clusters. Following screen shot represents output for K Means algorithm used with SMTP. It can be observed in the following screen shot that, each cluster contains some emails whose numbers are displayed against cluster id.

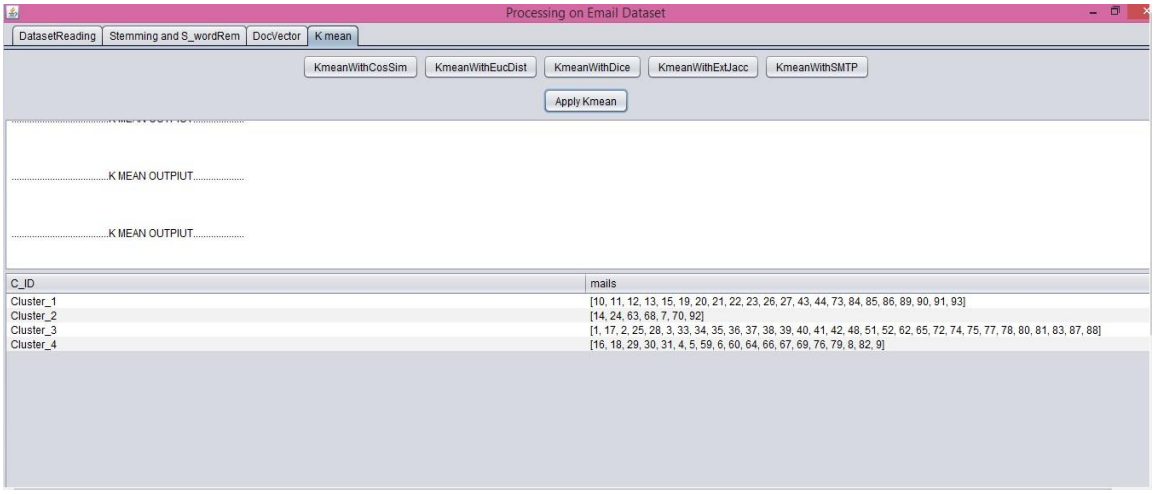

Fig. 2. Output of K means with SMTP

Accuracy is calculated using precision. Precision is defined as –(number of relevant record retrieved / number of irrelevant record retrieved+ number of relevant record retrieved) * 100.

The following table shows accuracy obtained with five similarity measures used with k means clustering algorithm.

Table 1.Accuracy for four clusters for five similarity measures using K Means algorithm

Similarity Measure	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cosine Similarity	36.1111	6.4516	20.8333	28.5714
Euclidean Dist	24.6575	25.00	2.38095	20.00
Dice Coefficient	23.3333	8.8235	23.9130	3.8461
Extended Jaccard Coefficient	24.2424	20.00	27.4509	26.3157
SMTP	53.8461	23.0769	83.3333	31.5789

It can be observed from above table, that the accuracy is more for SMTP. SMTP gives best results for all four clusters. Main difference between SMPT and other four similarity measures used is, SMTP considers presence or absence of words (features), while others consider frequency count i.e. number of times each term appears in given document. Advantages of SMTP can be discussed as follows. SMTP considers presence or absence of features than difference between two values associated with present feature. It also considers that similarity degree should increase when difference between two non-zero values of a specific term decreases. It also takes into account that similarity degree should decrease when the number of presence or absence of terms increases. SMTP takes into consideration one more important aspect that two documents are the least similar to each other if none of the terms have non-zero values in both documents. SMTP is symmetric similarity measure. The last and most important fact is that it considers standard deviation of term or feature which is taken into count for its contribution to similarity between the two documents.

The result analysis is also done for FCM algorithm applied on similarity matrix generated, for five distance measures- Euclidean distance, Cosine similarity, dice coefficient, extended Jaccard coefficient and SMTP. Again experts results are generated which compared with system generated results. Following screen shot represents output for FCM algorithm used with SMTP. Again for experimental purpose, 4 clusters are considered.

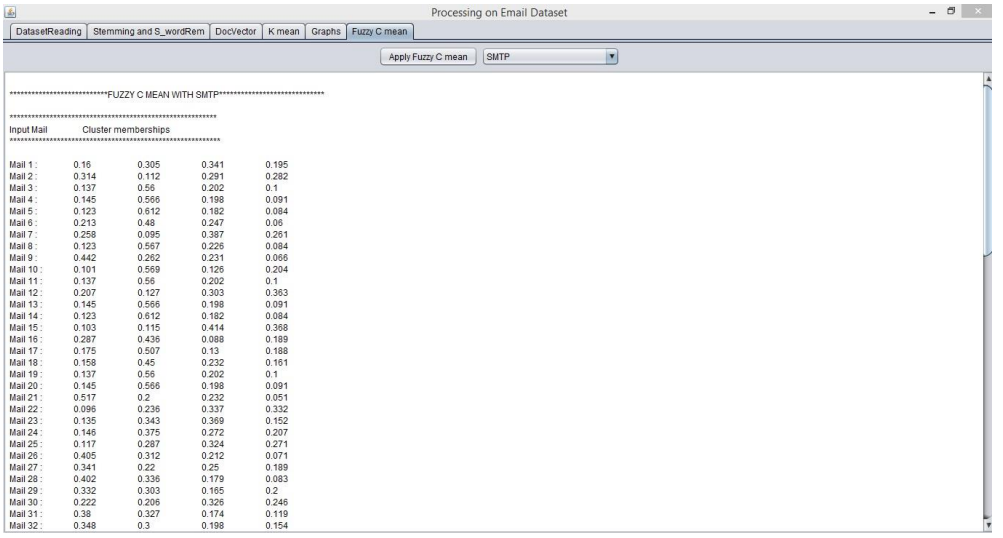

Fig. 3. Output of FCM with SMTP

In the above screen shot, membership of each email with each of the four clusters is displayed. Email is assigned to a cluster for which it has the highest membership. Accuracy is calculated using precision. Following table shows accuracy for FCM algorithm with five similarity measures.

Table 2. Accuracy for four clusters for five similarity measures using FCM algorithm

Similarity Measure	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cosine Similarity	37.1871	8.8765	22.76	29.8139
Euclidean Dist	25.7654	28.8712	5.00	21.2198
Dice Coefficient	27.1987	10.8876	24.90	4.00
Extended Jaccard Coefficient	25.4874	22.65	31.3451	28.98
SMTP	55.1238	24.00	86.387	35.87

As can be observed from table 2, FCM outperformed K Means in almost all four clusters. The results are represented in the following graph-

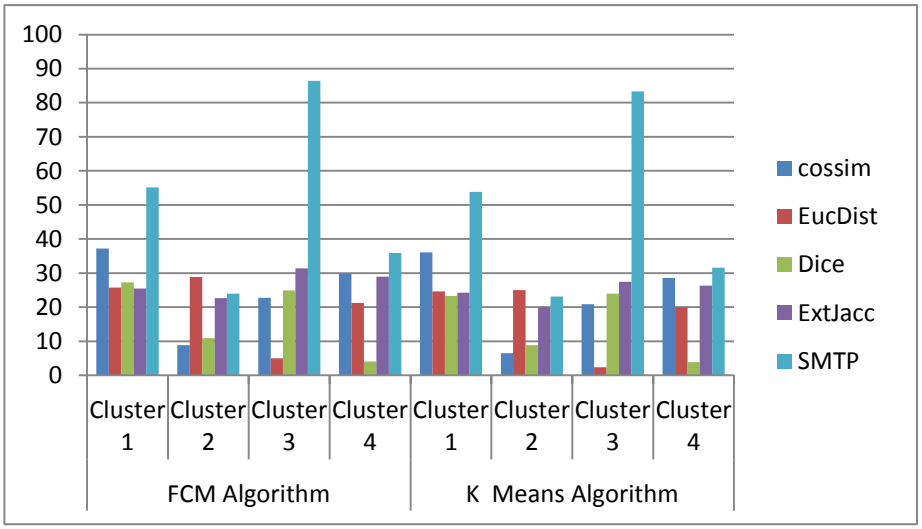

Fig. 4. Comparison between FCM algorithm and K Means algorithm with five distance measures

6. Conclusions

The main aim of FCM clustering algorithm is to minimize the objective function given in (1). The algorithm forms clusters by iteratively searching for a set of fuzzy clusters and the associated cluster centers that represent the structure of the data as best as possible. Both algorithms use distance measures. But FCM clustering uses distance measures along with fuzziness coefficients which control the degree of membership of each data item to a particular cluster. Also, because every document will have some membership value in each of the clusters, no useful document will ever be excluded from search results in case of fuzzy clustering; moreover fuzzy classification takes care of outliers in much better way. The clusters formed using FCM clustering is more accurate than clusters formed using k means clustering. The final conclusion is FCM clustering gives better results than k means clustering for given data set.

Acknowledgement

Inspiration and guidance are invaluable in every aspect of life, especially in the fields of academics, which I have received from my respected guide Prof. Y.V. Dongre. I would like to thank him for his endless contributions of time, effort, valuable guidance and encouragement she has given me. I also wish to thank everyone who has contributed to this work directly or indirectly.

References

[1] [Online] Availablehttps://www.cs.cmu.edu/~./enron/

[2] T. W. Schoenharl & G. Madey. Evaluation of measurement techniques for the validation of agent-based simulations against streaming data. Proc. ICCS 2008, Krakow, Poland.

[3] J. Han & M. Kamber. Data Mining Concepts and Techniques. 2nd ed. San Francisco ,CA, USA: Elsevier;2006.
[4] C.G. Gonzalez, W. Bonventi, Jr. & A.L.V. Rodrigues.Density of closed balls in real-valued and autometrized Boolean spaces for clustering applications. Proc. 19th Brazilizn Symp. Artif. Intel 2008; pp. 8-22.
[5] J. A. Aslam & M. Frost. An information-theoretic measure for document similarity. Proc. 26th SIGIR 2003; pp. 449-450.
[6] D. Lin. An information theoretic definition of similarity. Proc. 15th Int. Conf. Mach. Learn 1998 SanFrancisco, CA, USA.
[7] J.D'hondt, J. Vertommen, P.A. Verhaegen, D. Cattrysse & R.J. Duflou. Pairwise-adaptive dissimilarity measure for document clustering. Inf. Sci. 2010; Vol. 180, No. 12, pp. 2341-2358.
[8] R.W.Hamming. Error detecting and error correcting codes. Bell Syst. Tech. J. 1950; Vol. 29, No.2, pp.147-160.
[9] S. Kullback & R.A.Leibler. On information and sufficiency. Annu. Math. Statist. 1951; Vol. 22, No. 1, pp. 79-86.
[10] Yung-Shen Lin, Jung-Yi Jiang & Shie-Jue Lee. Similarity Measure for Text Classification and clustering. IEEE Transactions on Knowledge and Data Engineering 2014; Vol. 26, No. 7.
[11] Zadeh, L.A. Fuzzy sets. Information and Control **8** (3): 338–353 1965; doi:10.1016/s0019-9958(65)90241-x.
[12] Zadeh, L.A. Fuzzy Logic. Stanford Encyclopedia of Philosophy. Stanford University 2006.
[13] Zadeh, L. A. et al. Fuzzy Sets, Fuzzy Logic, Fuzzy Systems, World Scientific Press 1996; ISBN 981-02-2421-4
[14] Kosko, B. Fuzzy Thinking: The New Science of Fuzzy Logic.1994; Hyperion.
[15] Pratihar, D.K.: Soft Computing. Narosa Publishing House, New Delhi, India
[16] Dunn, J.C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well- Separated Clusters. J. Cybernet 1973; Vol. 3, pp. 32–57.
[17] Bezdek, J.C.Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers, Norwell, MA, USA, 1981.
[18] Pal, N.R.—Bezdek, J.C. .On Cluster Validity for the Fuzzy C-Means Model. IEEEFS 1995; Vol. 3, No. 3, p. 370.
[19] Albayrak, S.—Armasyali, F. Fuzzy C-Means Clustering on Medical Diagnostic System. Proc. Int. XII Turkish Symp 2003; on Artif. Intel. NN.
[20] Zhang, D.Q.—Chen, S.C. A Novel Kernelized Fuzzy C-Means Algorithm With Application in Medical Image Segmentation. Artif. Intel. Med 2004; Vol. 32, pp. 37–50.
[21] Migaly, S.—Abonyi, J.—Szeifert, F. Fuzzy Self-Organizing Map Based on Regularized Fuzzy C-Means Clustering. Advances in Soft Computing, Engineering Design and Manufacturing. J.M. Benitez, O. Cordon, F. Hoffmann, et al. (Eds.), Springer Engineering Series 2002; 2002, pp. 99–108.
[22] Sikka, K.—Sinha, N.—Singh, P.K.—Mishra, A.K. A Fully Automated Algorithm Under Modified FCM Framework for Improved Brain MR Image Segmentation. Magnetic Resonance Imaging. 2009 Vol. 27, No. 7, pp. 994–1004.
[23] Krinidis, S.—Chatzis, V. A Robust Fuzzy Local Information C-Means Clustering Algorithm.IEEE Trans. on Image Processing 2010; Vol. 19, No. 5, pp. 1328–1337.
[24] Belhassen, S.—Zaidi, H. A Novel Fuzzy C-Means Algorithm for Unsupervised Heterogeneous Tumor Quantification. PET. Medical Physics 2010; Vol. 37, No. 3, pp. 1309–1324.

Authors' Profiles

Ms. Deepa B. Patil is a post graduate student in Computer Engineering from Vishwakarma Institute of Information Technology under Savitribai Phule Pune University, Pune, Maharashtra State, India.

Prof. Yashwant V. Dongre is Assistant Professor in Vishwakarma Institute of Information Technology (VIIT) under Savitribai Phule Pune University, Pune, Maharashtra State, India. His area of interest includes database management, data mining and information retrieval. He has several journal papers to his credit published in prestigious journals.

How to cite this paper: Deepa B. Patil, Yashwant V. Dongre,"A Fuzzy Approach for Text Mining", International Journal of Mathematical Sciences and Computing(IJMSC), Vol.1, No.4, pp.34-43, 2015.DOI: 10.5815/ijmsc.2015.04.04