Proposed Method to Enhance Text Document Clustering Using Improved Fuzzy C Mean Algorithm with Named Entity Tag

Raghad M. Hadi*,M.Sc.(Lecturer) Soukaena H. Hashem**,Ph.D.(Asst.Prof.)
Abeer T. Maolood**,Ph.D.(Asst.Prof)

Abstract

Text document clustering denotes to the clustering of correlated text documents into groups for unsupervised document society, text data mining, and involuntary theme extraction. The most common document representation model is vector space model (VSM) which embodies a set of documents as vectors of vital terms, outmoded document clustering methods collection related documents lacking at all user contact. The proposed method in this paper is an attempt to discover how clustering might be better-quality with user direction by selecting features to separate documents. These features are the tag appear in documents, like Named Entity tag which denote to important information for cluster names in text, through introducing a design system for documents representation model which takes into account create combined features of named entity tag and use improvement Fuzzy clustering algorithms.

The proposed method is tested in two levels, first level uses only vector space model with traditional Fuzzy c mean, and the second level uses vector space model with combined features of named entity tag and use improvement fuzzy c mean algorithm, through uses a subset of Reuters 21578 datasets that contains 1150 documents of ten topics (150) document for each topic. The results show that using second level as clustering techniques for text documents clustering achieves good performance with an average categorization accuracy of 90%.

Keywords: Fuzzy clustering, documents datasets, information extraction, named entity.

*Al-Mustansiriyah University
**University of Technology

1. Introduction

Document clustering is an essential action in unverified document society, involuntary theme mining, charity to collection a regular of documents into clusters, by the impartial of exploiting intra cluster match and reducing inter cluster match [1] , where the progress there is no class labels delivered, as in document clustering, clustering can be complete in a semi supervised style where approximately related information is unified [2]. Document clustering involves the use of tags and tags extraction from documents header. Tags are groups of verses that label the fillings inside the cluster. In commonly document clustering is careful to be a consolidated procedure. Instances of document clustering contain e-mail message, blogs clustering for search users. The submission of document clustering canister considered to dual categories, connected and disconnected with internet. Connected submissions are frequently forced by competence difficulties once associated to disconnected submissions [3] Dimension decrease done through separated interested in piece choice and piece mining. Piece choice is the method for choosing reduced subgroups as of datasets and piece mining alters the great space of dataset to a new space with lower dimensions. The space reduction objective used for permit minor volume of data space for wider evaluations of the ideas checked in a text assembly. The most common technique for dimension reduction is Singular Value Decomposition (SVD) which is used to recognize forms in the relations among the terms and notions checked in a gathering of text. The SVD decreases the scopes by choosing sizes with top singular values. When a document term matrix A (m×n) is created, supposing that there are m documents and n terms. The SVD subtracts the document and term vectors by converting matrix A into three matrices U, S and V, which is specified by $A=USV^T$ [4].

Fuzzy C-Means is algorithm used for document clustering is an iterative process that updates the prototypes of the clusters defined initially from a fuzzy pseudo partition and the partition matrix giving the membership degree of each document to each cluster. This update tries to minimize the dissimilarity between a document and a cluster prototype. The pseudo-partition is defined as follows [6].

The Reuters 21578 datasets that one of a multi-dimensional datasets which container achieve organization jobs crossways unlike caring of groups. Aimed at, slightly organization might shape a dataset seeing

"Topics" classes somewhere "places" = "Canada" and consequently. There are documents going to 135 dissimilar classes of Topics. The shape an effectual classifier for 135 classes is also greatly exclusive in relations of labors to realize moral accurateness aimed at apiece class. Therefore, unknown nearby are not corporate restraints, is significant achieve a measurable examination to notice which are the greatest shared classes [7].

2. Related Works

The following related work will present document clustering techniques as much as related to the proposal:

- In 2015, A.Vijaya Kathiravan and P.Kalaiyarasi [8], proposed document clustering method as the method of cluster mechanically consortium documents into number of clusters. As a substitute of penetrating complete documents for relevant information, these clusters will improve the efficiency and avoid overlapping of contents. Birch hierarchical clustering algorithm that can be applied to any relational clustering problem, and its application to several non-sentence data sets has shown its performance to be comparable to k-means benchmarks allow patterns to belong to all the clusters with differing degrees of membership. The authors compared their work with hard clustering using birch algorithms. The result of comparing shows that their work avoids content overlap and able to achieve superior performance to k-means algorithms when externally evaluated on a challenging data set of famous quotations. In their proposed system, they used birch clustering algorithm that operates on cluster start with initial threshold and inserts points into the tree.
- In 2014, D. Renukadevi and S. Sumathi [9], proposed a method for developing of information technology and cumulative usability of internet. The authors shows that the extracted document is pre-processed, then the document is ranked using frequency of each word that can be calculated by using Term Frequency-Inverse Document Frequency method. After that the similar information is grouped together using fuzzy c-mean clustering which has been experimentally proved and verified by the results. Their proposed improved the clustering accuracy and it was less classification time.

- In 2015, Rashmi D thakare and Manisha R patil [10] proposed method to extract template from heterogeneous web documents using clustering. The authors shows that extraction from different web pages is studied which is done without any human data input. A template is well defined which would propose the framework to be used to describe how the values are inserted into the pages. The extraction algorithm is to extract values from web pages. This algorithm is trained to generate the template referring defined set of words having common occurrence. The authors implement the MDL principle to manage the unknown number of clusters then MinHash technique has been implemented to speed up the clustering process. The experimental results show that their work was effectively cluster web documents.
- In 2013, M.Gong.Y. Liang, W.Ma and J.Ma [11], presented a technique to enhanced FCM by put on the seed coldness amount toward the impartial purpose. The authors show the chief impression of seed approaches is to alter compound nonlinear difficulties in unique short-space piece to the simply solved difficulties in the great space. The additional method to contract through the limitation *m* is appreciating the organization of indecision happening the foundation of the fuzziness guide.
- In 2013, Yinghua Lu, Tinghuai Ma, and Changhong Yin [12], presented method to improve fuzzy c mean algorithm to deal with meteorological data on top of the traditional fuzzy c mean. In this paper the authors introduces the features and the mining process of the open source data mining platform WEKA. The experimental results show that their proposal was generated better clustering results than k-means algorithm and the traditional fuzzy c mean.
- In 2007, C. Hwang and F and C. H. Rhee [13], proposed the intermission form-2 fuzzy set into The Fuzzy clustering algorithm was combined with the intermission form-2 fuzzy set to achieve an indecision for fuzziness guide k. The authors compared their proposal with problem that need to specify and the result shows that the fuzzy clustering algorithm is informal near grow hit popular the resident modicums, although whatever that need to discovery is the worldwide dangerous.
- In 2015, P. Chiranjeevi, T. Supraja, and P. Srinivasa Rao [14], presented clustering as the task of grouping a set of objects in

such a way that objects in the same group are more similar to each other than to those in other groups . The authors shows that their suggested document clustering technique is practically useful document clustering method with high intra-similarity and low inter-similarity space using a brief survey on optimization method to text document clustering use of external source like wordnet.

3. The proposed method

In text mining tests the wide spread datasets used is the Reuters dataset, so, in this proposed method, the Reuters 21578 dataset is also used. The Reuters 21578 dataset contain of documents that performed on the Reuters Newswire in 1987. The dataset consists of 22 files: The first 21 files contain 1000 documents each, and the 22nd contains 578 documents. The formatting of the data is in SGML format. The original assortment then has lone 21,578 documents, and therefore is named the Reuters-21578 gathering. Which denote as "Reuters-21578, Distribution 1.0". Formerly generate a document term matrix which is just a matrix through documents as the rackets and terms by way of the pillars and a total of the occurrence of disputes as the jail cell of the matrix. The arrangement of the Reuters 21578 datasets show in table (1) [5].

Table (1): The composition of the Reuters datasets

Category Set	Number of Categories	Number of Categories w/ 1+Occurrences	Number of Categories w/20+ Occurrences
Exchanges	**39**	**32**	**7**
orgs	**56**	**32**	**9**
people	**267**	**114**	**15**
places	**175**	**147**	**60**
topics	**135**	**120**	**57**

The proposed method contains of dual stages: training stage and testing stage. The training stage objective is to adjust value of selected prototype vectors according to a set of documents di= {d1, d2, …, dn} which is training documents, each document corresponding its feature vectors (set of terms in each document approximately 1677 terms (features)), while the goal of the testing stage is to cluster the incoming documents into requirement clusters based on the prototype vectors produced from the

training stage. Figure (1) shows the block diagram of the proposed method

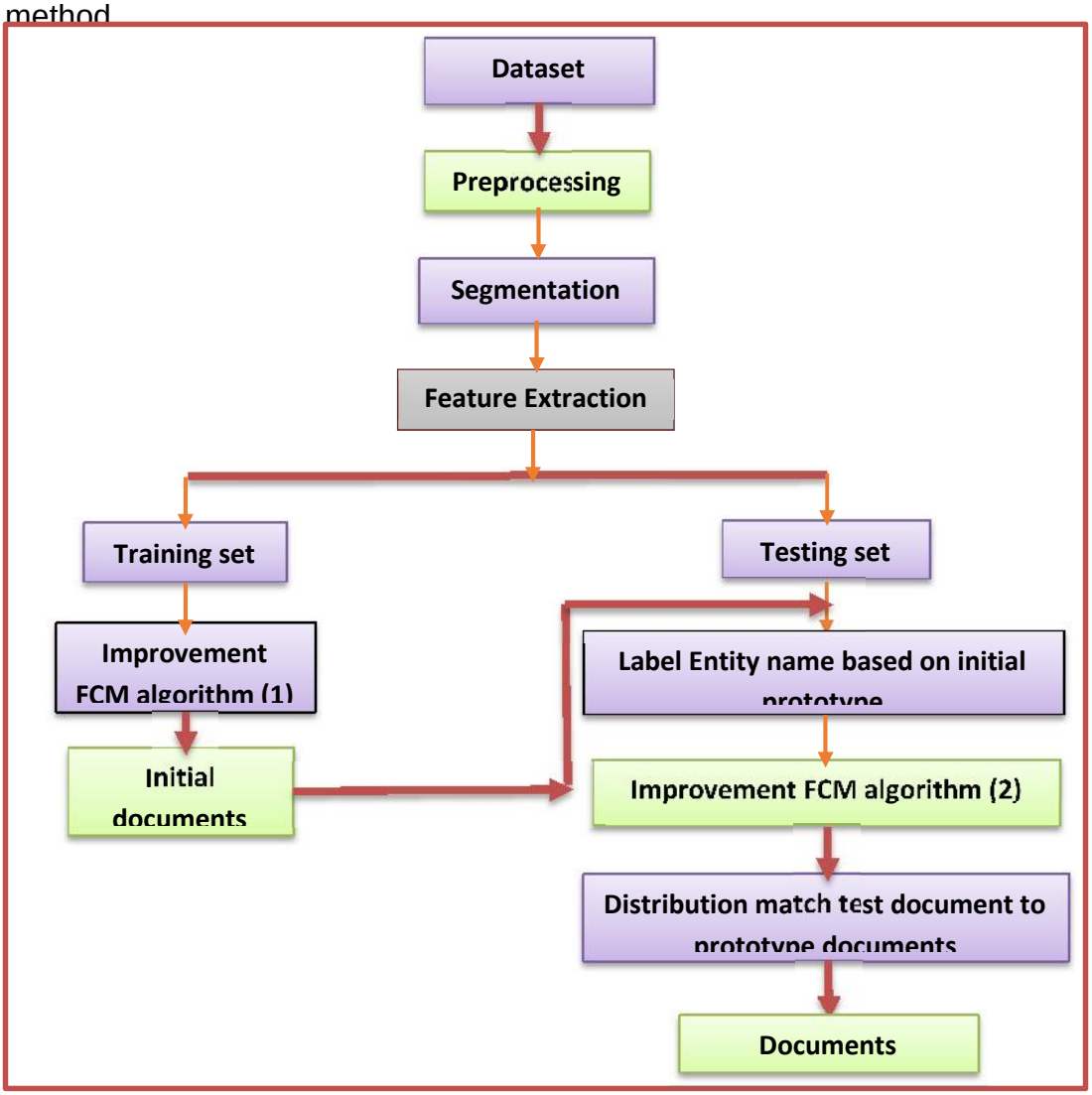

Figure 1:- Block diagram for the proposed system

3.1. Subset dataset and segmentation

The proposed method use the Reuters-21578 dataset, collect whole documents from datasets by split the documents and constructing two subsets. The documents are selected according around additional than unique tags sort otherwise unique document subject. Unique training subset involves the documents of single named entity tag, though the testing subset concluded together named entity and subject tag. A dataset initially, signified via training subset (TD1), encompasses 700 forms through single otherwise additional of the four tags Places, People, Orgs, and Exchanges, in which: 400 forms encompass individual one named entity tag each, 234 documents encompass dual named entity tags apiece, 55 documents have three named entity tags apiece, and 6 documents encompass four named entity tags apiece. The delivery for 700 documents through a four named entity tags is for example; places: 400 documents, people: 300 documents, organizations: 281 documents, exchanges: 86 documents.

A second dataset, signified by testing dataset (TD2), encompasses 450 documents through single or additional for four tags places, people, organizations and exchanges, in which: 400 documents encompass single named entity tag, 240 documents encompass dual named entity tags apiece, 60 documents have three named entity tags apiece and 6 documents encompass four named entity tags apiece. A delivery of the 700 documents across the four named entity tags is by way of: places: 400 documents, people: 300 documents, Organizations: 281 documents, exchanges: 86 documents.

3.2. Dataset preprocessing and feature extraction

collect whole documents for each category by using Body based feature, All body based features existing in the body of Reuter's document that includes: (body-keyword), (<body >), (body-java script) and etc. After these body, the content of document begin, each body document in datasets was represented using the bag-of-words approach, also these representation known as VSM. The proposed system added extended to the VSM which is Named Entity (NE) as following:

Named entity tag	TF-IDF for each document

Named entity tag represented by the following two value:

Category types	Document number in category

Terms Frequency (TF) = $\frac{\text{Number of times terms T appears in a document}}{\text{total number of terms in the document}}$ and the

Inverse Document Frequency (IDF) = $\log \frac{\text{total No.of document}}{\text{No.of document with term T appear in it}}$

For each term in document the proposed system calculate TF-IDF value. After extract features, the proposed method decomposition the TF-IDF matrix by using Singular Value Decomposition (SVD) in to three matrixes USV^T, then find k greatest chief scopes (through the top singular values in S matrix) is nominated and completely additional features stay absent. The summary matrix perfectly denotes the significant and dependable patterns underlying the data in TF-IDF matrix. The proposed method dropping the rank of the TF-IDF matrix incomes of eliminating unimportant info or clatter from the datasets it embodies.

3.3. Clustering with improvement fuzzy C mean algorithm

The proposed method use a technique to define document prototype that is based on firstly randomly select for prototype document [12] in each category which represent the number of cluster required then calculate the distance of all documents in cluster with selected document prototype and compare the distance with two threshold value, the number of documents whose distance is less than the given smaller threshold from the document prototype are ignore by the proposed system, and the document which distance is the biggest value from than given larger threshold it selected as the second document prototype. And so, repeat this method until there is no document have largest distance value from previous calculated value. When the proposed system applied this method and got the results it's ensure to escape the impartial function into local minima. The threshold value are definite through operators affording to the features of the datasets. The main steps of the training stage is presented in the algorithm (1) as following:

Algorithm 1: Training stage of proposal system using improvement FCM algorithm

Input:

- Documents datasets to be clustering.
- Number of clusters.
- Fuzziness parameter.
- Threshold $To_1 > To_2$
- Set iteration number, IT=1.
- Maximum iteration, maxi

Output:

- Document prototype vectors for each clusters (C_i).
  Membership matrix.

Procedures begin:

Step 1: Document datasets extraction, split document in to n category corresponded to user supervision at the selecting features based on the tags that appear in documents, entity named tags added to each document representation as following:

Named entity tag	Document contents

That's means:

Category name	Document number	Document contents

Then preprocessing the documents content, tokenization the document, remove the stop words and unwanted words, stemming the words and stored the pre-processed n-document as D_i , where i= 1,2,3…. N.

Step 2: Creation of document-term matrix and finding TF-IDF matrix of D_i , where T(terms) are created by counting the number of occurrences of each word produce by pre-processing step in each document , each column t_i show terms occurrence in each document D_i, then finding out the TF*IDF of D_i for each terms belong to it

$$TF=\frac{\text{Number of times terms T appears in a document}}{\text{total number of terms in the document}}$$ and

$$IDF = \log \frac{\text{total No.of document}}{\text{No.of document with term T appear in it}}$$

Step 3: store representation of each document for each category as the extended TF-IDF forms matrix representation as:

Named entity tag	TF-IDF for each document

Step 4: extraction the cluster centroid for each category by the following steps:

a) Put all documents which belong to one category into a set CC.
b) Remove one document from CC and put it in Centroid Set CS.
c) For each other document in CC, compute Euclidean distance from document in CC to document in CS :
   I. If distance < To_1, place document from CC in Centroid Set CS.
   II. If distance < To_2, remove document from CC.
d) Repeat from 2 until there are no more document in the set CC.
e) Each documents obtained in CS set is treated as the best documents prototype for one category which gives the best expected initial clusters centers, then pick for each category initial document prototype , any document from CS. Then the proposed system have best documents prototypes $c_1, c_2, \ldots\ldots\ c_i$, where i = number of required cluster.
f) compute cluster membership values U_{ik} as :

$$U_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{\|d_k - c_i\|}{\|d_k - c_j\|} \right)^{\frac{2}{m-1}}} \quad \ldots\ldots\ (1)$$

where $\|d_k - c_i\| = \left\| \begin{bmatrix} tf-idf_{1k} \\ tf-idf_{2k} \\ tf-idf_{3k} \\ \ldots\ \ldots\ \ldots \\ tf-idf_{Nk} \end{bmatrix} - \begin{bmatrix} c_{1i} \\ c_{2i} \\ c_{3i} \\ \ldots\ \ldots \\ c_{Ni} \end{bmatrix} \right\|$ …….. (2)

which represent the Euclidean distance between document $_{k,}$ and the document prototype vector i and

$$\|d_k - c_j\| = \left\| \begin{bmatrix} tf-idf_{1k} \\ tf-idf_{2k} \\ tf-idf_{3k} \\ \dots\dots\dots \\ tf-idf_{Nk} \end{bmatrix} - \begin{bmatrix} c_{1j} \\ c_{2j} \\ c_{3j} \\ \dots\dots \\ c_{Nj} \end{bmatrix} \right\| \dots\dots (3)$$

where **j**= {1, 2,…. C (number of cluster)}

which represent the Euclidean distance between document $_{k,}$ with all document prototype vector j where j = {1,2, …. , number of document clusters}.

Step 5: update document prototype vectors of the required clusters using:

$$C_j = \frac{\sum_{i=1}^{n}[U_{ij}]^m * d_i}{\sum_{i=1}^{n}[U_{ij}]^m} \dots\dots\dots\dots (4)$$

j (number of document clusters) = 1……. c.

i (number of document vectors) = 1…n.

U_{ij} = The degree of membership document i in the cluster j.

$$\mathbf{C_j} = \frac{U_{1j}{}^n \begin{bmatrix} tf-idf_{11} \\ tf-idf_{21} \\ tf-idf_{31} \\ \dots\dots \\ tf-idf_{N1} \end{bmatrix} + U_{2j}{}^n \begin{bmatrix} tf-idf_{12} \\ tf-idf_{22} \\ tf-idf_{32} \\ \dots\dots \\ tf-idf_{N2} \end{bmatrix} + \dots\dots\dots + U_{nj}{}^n \begin{bmatrix} tf-idf_{1j} \\ tf-idf_{2j} \\ tf-idf_{3j} \\ \dots\dots \\ tf-idf_{Nj} \end{bmatrix}}{U_{1j}{}^m + U_{2j}{}^m + U_{3j}{}^m + \dots\dots\dots + U_{nj}{}^m} \dots\dots (5)$$

Step 6: checking for stopping criteria, if Iteration number (IT) > max_i then stop, else increment iteration number, and go to step 3.

End algorithm (1).

The main steps of testing stage is shown in algorithm (2):

Algorithm 2: Testing stage of proposal system using improvement FCM algorithm

Input:

- Documents datasets, testing dataset (DT2) to be clustering.
- Number of clusters.
- Fuzziness parameter.
- Document prototype vectors from training stage.
- Set iteration number, IT=1.

Output:

- Clustering Document set.
- Membership matrix.

Procedures begin:

Step 1:

- For the document cluster centroid $C_1, C_2, \ldots, Ci$ from training stage and each input document $d_1, d_2, \ldots, d_k$, compute cluster membership values U_{ik} as :

$$U_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{\|d_k - c_i\|}{\|d_k - c_j\|} \right)^{\frac{2}{m-1}}} \quad \ldots\ldots\ldots\ldots (6)$$

Where

$$\|d_k - c_i\| = \left\| \begin{bmatrix} tf - idf_{1k} \\ tf - idf_{2k} \\ tf - idf_{3k} \\ \ldots \ldots \ldots \\ tf - idf_{Nk} \end{bmatrix} - \begin{bmatrix} c_{1i} \\ c_{2i} \\ c_{3i} \\ \ldots \ldots \\ c_{Ni} \end{bmatrix} \right\| \quad \ldots\ldots\ldots (7)$$

which represent the Euclidean distance between document $_{k,}$ and the document prototype Vector i. And

$$\|d_k - c_j\| = \left\| \begin{bmatrix} tf-idf_{1k} \\ tf-idf_{2k} \\ tf-idf_{3k} \\ \dots\ \dots\ \dots \\ tf-idf_{Nk} \end{bmatrix} - \begin{bmatrix} c_{1j} \\ c_{2j} \\ c_{3j} \\ \dots\ \dots \\ c_{Nj} \end{bmatrix} \right\| \dots \dots\dots (8)$$

where **j=** {1, 2, …. C (number of cluster)}, which represent the Euclidean distance between document $_{k,}$ with all document prototype vector j where j = {1, 2, …. , number of document clusters}.

Step 2: update document prototype vectors of the required clusters using:

$$C_j = \frac{\sum_{i=1}^{n}[U_{ij}]^m * d_i}{\sum_{i=1}^{n}[U_{ij}]^m} \dots\dots\dots (9)$$

j (number of document clusters) = 1…... c.

i (number of document vectors) = 1…n.

U_{ij} = The degree of membership document i in the cluster j.

$$\mathbf{C_j} = \frac{U_{1j}{}^m \begin{bmatrix} tf-idf_{11} \\ tf-idf_{21} \\ tf-idf_{31} \\ \dots\dots \\ tf-idf_{N1} \end{bmatrix} + U_{2j}{}^m \begin{bmatrix} tf-idf_{12} \\ tf-idf_{22} \\ tf-idf_{32} \\ \dots\dots \\ tf-idf_{N2} \end{bmatrix} + \dots\dots\dots + U_{nj}{}^m \begin{bmatrix} tf-idf_{1j} \\ tf-idf_{2j} \\ tf-idf_{3j} \\ \dots\dots \\ tf-idf_{Nj} \end{bmatrix}}{U_{1j}{}^m + U_{2j}{}^m + U_{3j}{}^m + \dots\dots\dots + U_{nj}{}^m} \dots\dots\dots (10)$$

Step 3: Assign label $C_1, C_2, \dots\dots, C_j$ to the tested document di, i= 1,2, …. n.

$$Dj = \begin{cases} c_1 & if\ U_{1j} > U_{nj} \\ c_2 & if\ U_{2j} > U_{nj} \\ \dots\ \dots\ \dots & \\ c_n & if\ U_{nj} > otherwise \end{cases} \dots\dots\dots\dots (11)$$

End algorithm (2).

4. Document Clustering Evaluation Methods

For evaluating the proposed document clustering approach, the proposed system perform two types of cluster evaluation; external evaluation, and internal evaluation. External evaluation is applied when the documents have tags. Internal evaluation is applied when documents tags are unknown.

4.1. External Evaluation

These measures are purity, entropy, and the F-measure. As the value of purity and F-measure increase it means that better clustering is achieved, on the other hand, as the value of entropy decreases it means better results are achieved.

- **Purity**

Purity is a measure for the degree at which each cluster contains single class label. To compute purity, for each cluster *j*, the number of occurrences for each class *i* are computed and select the maximum occurrence (max_{ij}), the purity is thus the summation of all maximum occurrences (max_{ij}) divided by the total number of documents *n*.

$$p = \frac{1}{n} \sum_{j}^{c} \max_{ij} \dots\dots\dots (12)$$

- **Entropy**

Entropy is a measure of uncertainty for evaluating clustering results. For each cluster *j* the entropy is calculated as follow

$$E(j) = \sum_{i=1}^{c} P_{ij} \log_2 \frac{1}{P_{ij}} \dots\dots\dots (13)$$

where, c is the number of classes, P_{ij} is the probability that member of cluster j belongs to class i,

$$P_{ij} = \frac{n_{ij}}{n_j} \dots\dots\dots (14)$$

where n_{ij} is the number of objects of class i belonging to cluster j, n_j is total number of objects in cluster j.

The total entropy E for all clusters is calculated as follow:

$$E = \sum_{j=1}^{k} \frac{n_j E(j)}{n} \dots\dots\dots (15)$$

Where k is the number of clusters, n_j is the total number of objects in cluster j, and n is the total number of all objects

- **F-measure**

F-measure is a measure for evaluating the quality for hierarchical clustering. F-measure is a mix of recall and accuracy. First the accuracy and recall are computed for each class i in each cluster j.

$$\text{Recall}(i,j) = \frac{n_{ij}}{n_i} \quad \ldots\ldots\ldots\ldots (16)$$

$$\text{accuracy}(I,j) = \frac{n_{ij}}{n_j} \quad \ldots\ldots\ldots\ldots (17)$$

The n_{ij} = number of documents of $class_i$ in cluster j, n_i = total number of document in $class_i$ and n_j = the total number of document in cluster j.
The F-measure of class i and cluster j is then computed as follow

$$\mathbf{F(i,j)} = \frac{(2*\mathbf{Recall(i,j)}*\mathbf{accuracy\ (i,j)})}{\mathbf{Recall(i,j)}+\mathbf{accuracy(i,j)}} \quad \ldots\ldots\ldots\ldots\ldots (18)$$

the maximum value of F-measure of each class is selected then, the total f-measure is calculated as following, where n is total number of documents, c is the total number of classes

$$F = \sum_{i=0}^{c} \frac{n_i}{n} \text{ Max } F(i,j) \quad \ldots\ldots (19)$$

4.2 Internal evaluation

For internal evaluation, the goal is to maximize the cosine similarity between each document and its associated center. Then, the results were divided by the total number of documents as following:

$$\text{Maximize} \frac{\sum_{j=1}^{k}\sum_{d=0}^{n_j}\cos(d_j,d)}{n} \quad \ldots\ldots\ldots\ldots\ldots\ldots (20)$$

where k denotes to number of clusters, n_j is the number of documents assigned to cluster$j$, d_j is the center of cluster.

5. Experimental Results

The proposed system usage the Reuters 21578 datasets for fuzzy clustering tests with number of documents selected for clustering are 1000 documents, actual number of classes 40. Table 2 shows the setting for the proposed system experiment.

Table 2: Setting for Experiment

Fuzzy C Mean parameters	Number of clusters	Set Randomly
	Fuzzier	Set Randomly
	Distance used	Euclidean distance
	Initial setting of membership weights	Randomly
	Stopping criteria	Stopping criteria < 0.005

Table 3 shows how the proposed system perform the external measures with including or not including entity names for VSM variety slight change. That one incomes that written customs only stand not important toward task of named entity tags in the Reuters 21578 datasets.

Table 3: External measures for fuzzy clustering in subset of Reuters 21578 with varied of C and fuzzy index m=2, threshold stop value=0.001

Purity	C=2	C=3	C=4	C=5	C=6
Named entity + documents features analysis	0.79	0.58	0.46	0.6	0.75
documents features analysis (TF-IDF matrix)	0.74	0.52	0.40	0.50	0.71
Entropy	**C=2**	**C=3**	**C=4**	**C=5**	**C=6**
Named entity + documents features analysis	1.50	1.13	0.90	1.13	1.22
documents features analysis (TF-IDF matrix)	1.56	1.15	0.95	1.2	1.5
F-measures	**C=2**	**C=3**	**C=4**	**C=5**	**C=6**
Named entity + documents features analysis	1.59	1.68	1.79	1.88	1.91
documents features analysis (TF-IDF matrix)	1.54	1.62	1.70	1.83	1.82

Table 4 as well as table 5 existing the external measures prices by diverse C and the threshold stop value (α) on the subset TD1 for the two models documents features analysis (TF-IDF matrix) and Named entity + documents features analysis, designed for apiece rate of C present is an best value of α giving the best clustering quality.

Table 4: The Purity measures with varied C and α on subset TD1 for Named entity + documents features analysis

Purity	α= 0.1	α= 0.2	α = 0.3	α =0.4
C= 2 (with single class label)	78300	105	566	42.4
C=3 (with single class label)	166.9	6514	1464	726
C=4 (with single class label)	169.8	76.6	3037	807
C=5 (with single class label)	168.5	93.4	38.3	234

Table 5: The Purity measures through diverse C and α on subset TD1 for documents features analysis only

Purity	α= 0.1	α= 0.2	α = 0.3	α =0.4
C= 2 (with single class label)	34300	99	333	26.1
C=3 (with single class label)	122.9	3514	1022	390
C=4 (with single class label)	125.8	33.6	1032	432
C=5 (with single class label)	124.5	53.4	13.6	182

6. Conclusion and Future work

The proposed method applied firstly on the outmoded fuzzy clustering algorithm participate it into Reuters 21578 datasets, related with progress the traditional fuzzy c mean in stretch of the choice of original cluster centers. Secondly the proposed method assume a new method to determine cluster centers, then it is generous the greatest marks for evaluation measures entropy and f-measure which are standard external measures and are additional significant to justice legitimacy of document clusters. The results show that using second level as clustering techniques for text documents clustering achieves good performance with an average categorization accuracy of 90%.

In the future research, the proposed method can improve the performance of the FCM in the field of another datasets from other aspects.

References

[1] Han, Kamber, “Data Mining Concepts and Techniques”, Morgan Kauffman Publishers, 2014.

[2] Stuti Karol, Veenu Mangat, “Evaluation of text document clustering approach based on particle swarm optimization”, 2013.

[3] Pengtao Xie, and Eric P.Xing, “Integrating Document Clustering and Topic Modeling”, Tsinghua University, China, 2014.

[4] P. Chiranjeevi, T. Supraja and P. Srinivasa Rao, “A Survey on Extension Techniques for Text Document Clustering”, International Journal of Advanced Research in Computer Science and Software Engineering, 2015**.**

[5] Ian Kloo,”Textmining: Clustering, Topic Modeling, and Classification”, August 2015.

[6] Anita Krishnakumar, “Text categorization building a KNN classifier for the Reuters-21578 collection”, Department of Computer Science, 2006.

[7] Tanagra Data Mining Ricco Rakotomalala, “Text mining” with Knime and RapidMiner. Reuters Text Categorization, 2016.

[8] Dr.A.Vijaya Kathiravan and P.Kalaiyarasi, “Sentence-Similarity Based Document Clustering Using Birch Algorithm”, 2015.

[9] D. Renukadevi, and S. Sumathi, “Term Based Similarity Measure for text classification and clustering using fuzzy c mean algorithm”, International Journal of Science, Engineering and Technology Research (IJSETR), 2014.

[10] Rashmi D thakare and Manisha R patil, “Extraction of Template using Clustering from Heterogeneous Web Documents”, International Journal of Computer Applications 2015.

[11] M.Gong.Y. Liang,W.Ma ,“Transactions on image processing”, 2013.

[12] Yinghua Lu, Tinghuai Ma and Changhong Yin, “Implementation of fuzzy c mean clustering algorithm in meteorological data”, 2013.

[13] C. Hwang and F. C. H. Rhee, J. “Transactions on Fuzzy Systems”, 2007.

[14] P. Chiranjeevi, T. Supraja, P. Srinivasa Rao, “A Survey on Extension Techniques for Text Document Clustering”, 2015

طريقة مقترحة لتحسين عنقدة الوثائق النصية باستخدام خوارزمية العنقدة المضببة المحسنة مع علامات اسماء الكيانات

م. رغد محمد هادي * أم.د.سكينة حسن هاشم ** أم.د.عبير طارق محمود **

المستخلص

عنقدة الوثائق النصية يعني تجميع الوثائق والنصوص المتشابهة الى عناقيد وهذا التجميع للوثائق غير خاضع للرقابة ، عند استخراج البيانات المهمة من النص وأستخراج موضوع غير الطوعي. النموذج الأكثر شيوعا لتمثيل الوثائق هو نموذج متجه الفضاء (VSM) الذي يجسد مجموعة من الكلمات المهمة الموجودة في الوثائق ، والاساليب القديمة في تجميع الوثائق المتعلقة كانت تفتقر الى اتصال المستخدم. النظام المقترح في هذا البحث حاول أكتشاف كيفية تجميع هذه الوثائق كي تعطي جودة أفضل مع تدخل المستخدم عن طريق تحديد ملامح لفصل هذه الوثائق. هذه الميزات تظهر كالعلامات في الوثائق، مثل علامات الكيان المسماة التي تدل على معلومات هامة عن أسماء تستخدم للتصنيف في النص، من خلال تصميم نظام يستخدم لتمثيل الوثائق والذي يأخذ في نظر الاعتبار إنشاء نموذج الفضاء ناقلات (VSM) مع ميزات مشتركة من كيان مسمى بالعلامات ويستخدم خوارزمية تحسين العنقدة المضببة. تم اختبار النظام في مستويين، يستخدم المستوى الأول الوحيد VSM مع FCM التقليدي، ويستخدم المستوى الثاني VSM مع ميزات مشتركة من ا لكيان مسمى مع استخدام تحسين FCM الخوارزمية، من خلال استخدام مجموعة فرعية من بيانات رويترز 21578 قاعدة البيانات التي تحتوي على 1150 وثيقة متكونة من عشرة مواضيع و(150) وثيقة لكل موضوع. وأظهرت النتائج أن استخدام المستوى الثاني قد حقق أداء جيدا مع متوسط دقة تصنيف 90٪ مقارنة مع ثقنيات تجميع الوثائق النصية الاخرى.

*الجامعة المستنصرية

**الجامعة التكنولوجية