Journal of College of Education for pure sciences(JCEPS)
Web Site: http://eps.utq.edu.iq/ Email: com@eps.utq.edu.iq
Volume 7, Number 3, September 2017

Data Clustering Using Fuzzy Approach

Raghad M. Hadi , Dr. Soukaena H. Hashem , Dr. Abeer T. Maolood

raghad_alrudieny@yahoo.com , soukana.hassen@yahoo.com , abeer282003@yahoo.com

University of Technology - Department of Computer Science

Abstract

In largeComputers;the huge volume of files actually generate disorder to analyze it. So, it desiresto design a clustering techniques which reduce the costs of analysts. Document clustering is an essentialprocess in text mining, which retrieve the information with an acceptable accuracy, which can be achieved by fuzzy clustering.

Reuters 21578 dataset is used for experimental purpose, the proposed system was tested by using Reuters 21578 datasets according to the time required to cluster data. The proposed system improves data clustering algorithms by construct required fuzzy clusters. The proposed system showed a good result compared with clustering techniques in comparing with other clustering techniques in time efficiency.

Keywords: Data mining, clustering,FCM algorithm.

Journal of College of Education for pure sciences(JCEPS)
Web Site: http://eps.utq.edu.iq/ Email: com@eps.utq.edu.iq
Volume 7, Number 3, September 2017

تجميع مجموعة من الوثائق باستخدام التصنيف المضبب

م. رغد محمد هادي ، أ.د.م. سكينة حسن هاشم ، أ.د.م. عبير طارق محمود
abeer282003@yahoo.com , soukana.hassen@yahoo.com, raghad_alrudieny@yahoo.com
الجامعة التكنلوجية - قسم علوم الحاسبات

الخلاصة

في أجهزة الكمبيوتر هناككم ه ائل من الملفات والتي تسبب اضطراب في تحليلها. لذلكتم تصميم نظاملتجميع الوثائق وقام هذا النظام بتقليل الجهد المطلوب في تحليل هذه البيانات. تصنيف او تجميع البيانات هي من العمليات الاساسية في تنقيب البيانات والتي تقوم باسترجاع المعلومات من هذه الوثائق بدقة مقبولة والتي تم تحقيقها باستخدام خوارزمية التصنيف المضبب . تم استخدم رويترز 21578 كبيانات كبيرة لأغراض التجربة. وتم أختبار النظام على هذه البيانات مقارنة بالوقت المطلوب لتصنيفها .النظام المقترح اعطى نتائج جيدة مقارنةبتقنيات التصنيف السابقة. وهذه المقارنة كانت على كفاءة الوقت المستغرق باجراء التصنيف.

الكلمات المفتاحية: تنقيب البيانات ، التصنيف، خوارزمية التصنيف المضبب.

1. Introduction

Computers haveexcessivesignificance in the world of technology. Theprogress in computer architecture effects in processing controlplusvast storage space in computers (which can haveactualmassive amount of data in it). But this maylead to create problem if somebodyneeds to examine anexact file in it [1],document clustering is used this problem.Document clustering (referred to as text clustering) is one of the greatestchiefin writing mining, the assistance the formingtobigquantity of documents into clusters. These tools have been developed to help computer documentsto be organized [2]. But such computer detained devices holds enormous regular of files and documents like huge Reuters 21478 datasets, thus it is not relaxed to do the study of all and each files independently[3].

2. Related works

The approach presented in [9] showed that different users may have different search goals , and in [9] went to improve search goal by analyzing search engine query , determineunlikeworker search aims for a request by clustering the projectedcommentmeetings and using the pseudo-documents to better represent the feedback sessions for clustering using Fuzzy C Means. The fuzzy comparison createdidentity- building algorithm which is used and with a new optimization way to plot responsemeetings to virtualdocuments that can capably imitate workerdatarequests. In [10] offerings a smallimpression of approaches for fuzzy gathering and conditionswantedgoodsaimed atabest fuzzy document clustering algorithm. Founded on these standards we selectedunique of the fuzzy clustering greatestprotrudingmeans, additionalexactly probabilistic c-means. In [11] used improved FCM for image processing technique, segmentation linked, an improved FCM combining mean shift algorithm is proposed to improve the segmentation pictorial effects and competence of traditional FCM [11].

3. Background ofPreprocessing and Clustering Algorithm

3.1 Removal of Stop word

The dataset must be Pre-processing as ended to eliminate the break words and stem words which are measured in the fewer significant to increase value and efficacy of facts. Severalcast-offverses in English are unusable in Information Retrieval (IR) and writing mining [4]. These words arenamedas 'Stop words'. Stop-words, which are linguistic-exactusefulverses, are commonverses that bringnot at allevidence (i.e., pronouns, prepositions, conjunctions). Samples of these words contain 'the', 'of',' and', 'to', etc. These break words are acquirekept in the record. Dataset is overloaded in to alternativerecord. Now stop words in dataset is impassive by relating with the stop word record[5].

3.2 Stems

These algorithm used in ordertodecreaseeveryverses with itssimilar stem to a publicmethod,in the information-retrieval it's suitableto use stemsin numerousparts of the work.Scientists in numerousparts of computational linguistics and information retrieval discoverthe wanted step, thenaimed tochangedetails. In mechanical morphological examination, the origin of a termmight be of fewerdirectattention than his suffixes.The method to stemsoccupiednowcomprises a dualpoint stemming scheme. The initialstage is the stemming processcorrect, regains the trunk of a term by eliminatingthelong-lastinglikelyfinalethatcompetitionsunique on a slantstowed in the workstation. The second stageholder'sequal'sterm with unequallist," typicallyexamplesis a "similar" stem differssomewhat in meaninggiving to the suffixes initiallytrailed it [6].

3.3 Vector Space Model

The vector space model is the method which all documents required to be clustering must be represented to this model,and can be identify it as the common model which represents a set of documents as vectors in a common vector space.In theclassic method, the documents d_i is measuredas vector, $\mathbf{d_i}$, in the word-universe (regular of "verses"). In its humblestmethod, separately document is denotedthrough the (TF) vector,$\mathbf{d}$tf= (tf_1, tf_2… tf_n), wherever$tf_i$is the occurrence of the *ith*term in the document.

$$tf_i = \frac{\text{Number of times terms T appears in a document}}{\text{total number of terms in the document}} \ldots. (1)$$

In calculation, theusage of this perfect that massesindividuallywordconstructed on its opposite document Occurrence(IDF) in the documentcollection.

$$\text{IDF}= \log_e \frac{\text{total No.of document}}{\text{No.of document with term T appear in it}} \ldots..(2)\ .$$

tf_i(term frequency) thatquantityin what wayrepeatedly a wordseems in a document, IDF (Inverse Document Frequency) thatquantityin what waysignificant a word is. TF-IDF= tf_i *IDF [7].

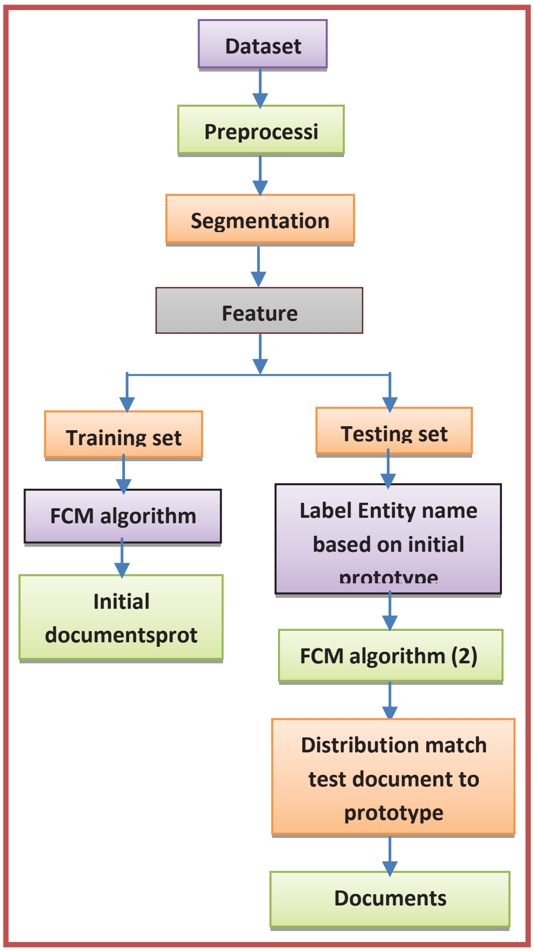
3.4 Clustering

A group (cluster) is clearby way of a subgroup of data items of the dataset that have its placecollected. The outcome of a fuzzy gatheringprocess is a fuzzy separating of datasets.

Clustering is a manner of separating a regular of informations (or items) obsessed by a usual of significantsubstitute-classes, termed clusters. Clustering assistancesworker to recognize the normalcombination or assembly in a dataset. It castoffalsoby way of a stance-onlyinstrument to growvision into data deliveryotherwise as a pre-processing stageaimed atadditionalprocedures. A respectable clustering techniqueresolvefoodtallexcellence clusters nowthat the intra-class resemblance is great. And the inter-class resemblance is little. The excellence of a clustering outcometoobe contingent on together the resemblancequantitycast-offvia the technique and its application [8].

4. Proposed Work

The proposed algorithm uses the techniques for Document Clustering to facilitate the large datasets analysts to do their work efficiently. The flowchart of the algorithm described here with the following steps of the proposed techniques:



4.1 Data Collection

The first main part of the proposed algorithmis assembles Reuters 21578 datasets which is available as training document set and testing documents set for text mining. The header were categorizedautomatically by Reuter'sworkers. Labels fit in to 5 dissimilargroup classes, such as 'people', 'places' and 'topics'. The total number of classes is 672, nonethelessseveral of them happenlone very seldom. Some documents fit to severalunlikeclasses, others to single one, and some have no class. Buthereneedseveralhard work to fresh the storeof such datasets, and increase it for usage in precise research. The currentfolder of these datasets is divided in 22 files of 1000 documents delimited by SGML tags, and from these files the documents dropped into 9603 training documents and 3299 testing documents and 8676 unused documents, it takes about 27 MB.

The categories in this dataset come from five classes:

- Exchanges: financial exchanges, e.g., "nasdaq"
- Organizations: named entities of organizations, e.g., "GE"
- People: named entities of people, e.g. "Paul Volcker"
- Places: named entities of places, e.g., "Australia"
- Topics: economic subject categories, e.g., "coconut", "gold", "money supply"

4.2 Preprocessing

The second main part of the proposed system is preprocessing part. To provide the proposed system with only the required data, so it's necessary to clean text documents by the pre-processing step of the proposed algorithm. The pre-processing step used in our proposed algorithm is described below with figure (2):

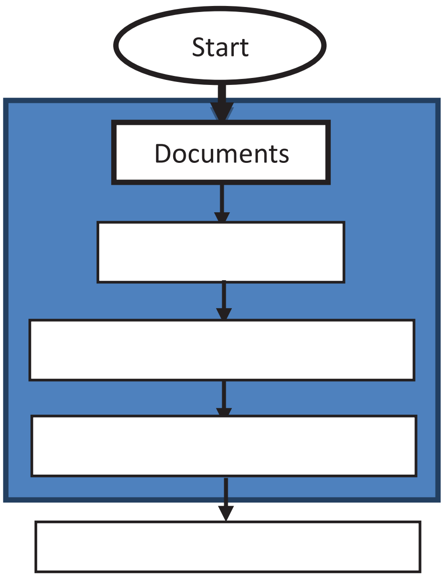


Figure 2: Preprocessing process

1) **Removal of Stop Words**

The proposed algorithm preserved a stop word dictionary having all possible stop words. By compared the words of the documents text with in the words store in stop word dictionary if found remove it. As well as the proposed system use the stop word creation methods moved by Zipf‟s law, including: delete the word that shows in the input text once (occur once), i.e. singleton words (TF1). And consider removing words with low (TF-IDF) value by first remove stop words from word vector using stop words list.

2) **Stemming**

The proposed system use porter stemming algorithm with enhancement on its rules, at each step, a certain suffix is deleted by uses of set rules. These rules are substitution rule which is applied when a set of conditions match to this rule so to reduce number of words, to have exactly matching stems, and to save memory space and time. The proposed system used Porters algorithm and table look up approach by having two dictionaries, one for various irregular English words, and another for various suffixes. To applied the following: Root = past simple or past participle.

Suffixed = root + suffix.

3) Create tf-idf

The documents denoted to it with (di)wasmeasuredas vector, di in the wordspace (list of terms), in its modestmethod, individually document is embodiedvia the TF vector, $d_{tf} = \{tf_1, tf_2, \ldots\ldots, tf_n\}$, Where$tf_i$ is the occurrence termi in the document.

$$tf_i = \frac{\text{Number of times terms I appears in a document}}{\text{total number of terms in the document}} \ldots (1)$$

Furthermore, in this model the terms areencumbrancesfounded on its inverse document Frequency (IDF) in the document gathering.

$$\text{IDF} = \log_e \frac{\text{total No.of document}}{\text{No.of document with term T appear in it}} \ldots.(2)\ .$$

tf_iquantityin what way a term looks in a document, IDF quantityin what waysignificant a term is. TF-IDF=tf_i *IDF, in table 1 show the frequency of each terms in datasets and TF value with IDF value.

Terms	TF value	IDF value	TF-IDF value
week	0.0108	4.3027	0.0464
behia	0.0144	8.0163	0.1153
cocoa	0.0216	7.3232	0.1581
come	0.0072	6.9177	0.0498
tempora	0.0072	8.0163	0.0577
have	0.0072	3.7536	0.0270
commissaria	0.0180	8.0163	0.1442
said	0.0180	1.7174	0.0309
period	0.0072	5.9369	0.0427
year	0.0072	2.9226	0.0210
arrive	0.0072	8.0163	0.0577

februari	0.0108	4.8383	0.0522
bag	0.0180	6.9177	0.1244
kilo	0.0072	6.9177	0.0498
total	0.0108	4.7582	0.0513
against	0.0108	5.1831	0.0559
consign	0.0072	8.0163	0.0577
still	0.0108	6.4069	0.0691
crop	0.0180	6.6300	0.1192
export	0.0072	4.3528	0.0313
dlr	0.0504	2.5191	0.1269
port	0.0108	6.4069	0.0691
open	0.0072	6.2246	0.0448
north	0.0476	6.6300	0.3157
texa	0.0588	6.4069	0.3769

4.3 Clustering

The third main part of the proposed algorithm is clustering the set of document using FCM algorithm. Figure 3 shows theGeneral layout of FCM-Document clustering.

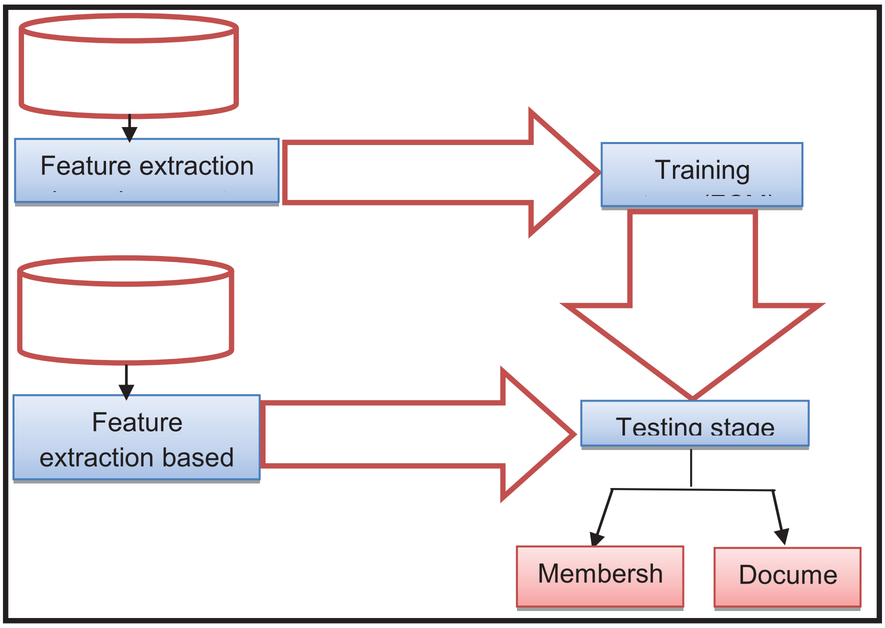


Figure 3: General layout of FCM-Document

The extraction feature set can be used by the FCM to define the prototype of each cluster, i.e., C= $\{C_1, C_2, \ldots. C_n\}$. The proposed FCM contains of dual stages: training stage and testing stage. the training stage objective is to adjust value of prototype vectors according to a set $d_i = \{d_1, d_2, \ldots, d_n\}$ of 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.

4.3.1 Training stage

The training stage objective is to construct initial document cluster center, and the number of clusters are determine by the user. The value of cluster number must be greater than 2 and less than 6.The main steps of the training stage are presented in algorithm (1), firstly, the p random document prototype vectors are selects to represent the initial centers for the training document datasets this datasets constructed on the two datasets DT1 and DT2 using the DT1 in the training stage to constructed the initial documentcenters because Fuzzy c mean relies on theinitial cluster centers and on initial membership degrees of all document in DT1 to projected clusters.

Algorithm 1: Proposal Training stage of FCM

Input:

- Documents datasets to be clustering.
- Number of clusters.
- Fuzziness parameter.
- Initialize randomly document prototype vectors.
- Set iteration number, IT=1.
- Maximum iteration, maxi.

Output:

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

Step 1: Document datasets extraction and preprocessing, submit the proposed system to extract the document from datasets , tokenization the document , remove the stopwords 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= $\frac{\quad}{\quad}$.

Step 3: Extraction the cluster centroid by the following steps:

- For the c clusters $C_1, C_2, \ldots\ldots, C_i$, and each documents $,d_1,d_2,\ldots, d_k$, c_i, and c_j is document prototype vectors, compute cluster membership values U_{ik} as :

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

$$\|d_k - c_i\| = \left\|\begin{bmatrix} tf-idf_{1k} \\ tf-idf_{2k} \\ tf-idf_{3k} \\ \ldots\ldots\ldots \\ tf-idf_{Nk} \end{bmatrix} - \begin{bmatrix} c_{1i} \\ c_{2i} \\ c_{3i} \\ \ldots\ldots \\ c_{Ni} \end{bmatrix}\right\| \ldots \quad (4)$$

Which represent the Euclidean distance between document_k, andthe document prototype vector i. And

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

Which represent the Euclidean distance between document_k, with all document prototype vector j where j = {1,2, …. , number of document clusters}

Step 4:update document prototype vectors of the required clusters using:

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

$$C_j = \frac{U_{1j}{}^m \begin{bmatrix} tf-idf_{11} \\ tf-idf_{21} \\ tf-idf_{31} \\ \ldots\ldots \\ tf-idf_{N1} \end{bmatrix} + U_{2j}{}^m \begin{bmatrix} tf-idf_{12} \\ tf-idf_{22} \\ tf-idf_{32} \\ \ldots\ldots \\ tf-idf_{N2} \end{bmatrix} + \cdots\ldots\ldots + U_{nj}{}^m \begin{bmatrix} tf-idf_{1j} \\ tf-idf_{2j} \\ tf-idf_{3j} \\ \ldots\ldots \\ tf-idf_{Nj} \end{bmatrix}}{U_{1j}{}^m + U_{2j}{}^m + U_{3j}{}^m + \cdots\ldots\ldots + U_{nj}{}^m} \ldots. (7)$$

Step 5:checking for stopping criteria, if IT > maxi then stop, else increment iteration number, and go to step 3.

End.

4.3.2. Testing stage

The testing stage objective is to cluster the new documents from DT2 dataset depended on the output from the training stage which represented bythe calculateddocument centroid from training stageand used it as input to testing stage. Algorithm (2)demonstrates the steps of testing stage of Fuzzy c mean to documents clustering.

Algorithm 1: Proposal Testing of FCM

Input:

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

Output:

- Document clusters (C_i) and Membership matrix.

Journal of College of Education for pure sciences(JCEPS)
Web Site: http://eps.utq.edu.iq/ Email: com@eps.utq.edu.iq
Volume 7, Number 3, September 2017

Step 1:

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

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

$$\text{Where } \|d_k - c_i\| = \left\|\begin{bmatrix} tf-idf_{1k} \\ tf-idf_{2k} \\ tf-idf_{3k} \\ \cdots\cdots\cdots \\ tf-idf_{Nk} \end{bmatrix} - \begin{bmatrix} c_{1i} \\ c_{2i} \\ c_{3i} \\ \cdots\cdots \\ c_{Ni} \end{bmatrix}\right\| \ldots (9)$$

Which represent the Euclidean distance between document_k, andthe document prototype vector i.

And

„ḆB F Ḛ@Ḇ sG _ „? sF _
ḆB F Ḛ@Ḇ tG ? tF
.@G F ? F . L q ḆB F Ḛ@Ḇ uG F ? uF q S DANA
ˇḆB F Ḛ@Ḇ 0 G ˇ? 0 F

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:

% F L ^J ELs >7 EF ? I @ E / ^J ELs >7 EF ? I

„ F _ „ F _ „ F _
F F F
F E F Efi E F

C_j= ˇ F z ˇ F z ˇ F z / 7 sF I E 7 tF I E 7 uF I Efi E 7 JF I :st;

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

Table 2			
Number of document (samples)	**Time (second)**		
	Pre-processing	**Clustering**	**total**
10	14.45	1.523	15.973
25	25.45	3.456	28.90
50	40.134	6.854	46.988
100	48.63	12.62	61.25

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:

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

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

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

$$Dj = \begin{cases} c_1 \;\; if\ U_{1j} > U_{nj} \\ c_2 \;\; if\ U_{2j} > U_{nj} \\ \ldots\ldots\ldots \\ c_n \;\; if\ U_{nj} > otherwise \end{cases}$$

End.

5. Experimental Results

Theproposed system used the Reuters 21578 datasets for fuzzy clustering tests with number of documents selected for clustering are 1000 documents, actual number of classes 40. Table 3 shows the setting for the proposed system experiment. Theoutput of the proposed system is a quantityof N clusters and for individuallydocument vector a set of numbers that denoteto themark of membership in each cluster. And the proposed system initiatesits work by the pre-processing step and its time takings for cluster different documents from the Reuters 21578is looks better, to calculate the processing time of clustering these document, the results shownin agiven table(table 2).

Table 3: 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< .005

Table 4 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.

5.1.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.

$$\mathbf{p = \frac{1}{n}\sum_{j}^{c} max_{ij}} \dots\dots\dots (13)$$

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. Conclusions

The proposed method applied firstly on the outmoded fuzzy clustering algorithm participate it into Reuters 21578 datasets, related with progress the fuzzy c mean in stretch of the choice of original cluster centers. Then it is generous the greatest marks for evaluation measuresPurity which it additional significant to justice legitimacy of document clusters. The results show that using fuzzy c mean algorithm 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.

7. References

1- Luís 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, January 2013.

2- U.S. Department of Justice, "Electronic Crime Scene Investigation: A Guide for First Responders", I Edition, NCJ 219941, 2008, http://www.ncjrs.gov/pdffiles1/nij/219941.

3- Vilas V Pichad, Sachin N Deshmukh, "Elevating Document Clustering For Forensic Analysis Investigation System", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 7, July 2015.

4- Keharsingh, Dimple Malik, Naveen Sharma "Evolving Limitations In K-means Algorithm In Data Mining And Their Removal", IJCEM International Journal of Computational Engineering And Management Volume 12, April 2011.

5- PriteshVora, BhaveshOza "A Survey On K-mean Clustering And Particle Swarm Optimization" International Journal Of Science And Morden Engineering (IJISME) ISSN: 2319-6386, Volume 1, issue-3, Feb 2013.

6- Julie Beth Lovins, "Development of a Stemming Algorithm", Massachusetts Institute of Technology, Massachusetts 02139, Mechanical Translation and Computational Linguistics, vol.11, nos.1 and 2, March and June 2011.

7- Michael Steinbach George KarypisVipin Kumar, "A Comparison of Document Clustering Techniques", Technical Report 00-034 ,2014.

8- B. Vidhya and R. PriyaVaijayanthi, " Enhancing Digital Forensic Analysis through clustering ", international journal of innovative research, Document in Computer and Communication Engineering, March 2014.

9- N.Vidhyapriya, S.Sampath, "Fuzzy C Means Algorithm for inferring User Search Goals with Feedback Sessions", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 1, January 2015.

10- MatjazJursic, Nada Lavrac, Jozef Stefan Institute, "Fuzzy clustering of documents", 2013

11- BingquanHuo and Fengling Yin Binzhou Polytechnic, Shandong, China huobingquan3@126.com, "Image Segmentation Using Mean Shift Based Fuzzy C-Means Clustering Algorithm: A Novel Approach ", International Journal of Multimedia and Ubiquitous Engineering Vol.10, No.5 (2015).