# طريقة مقترحة لتحسين خوارزمية العنقدة المضببة

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

الجامعة التكنلوجية قسم علوم الحاسوب

**المستخلص**

أن تحسين خوارزمية العنقدة المضبب قد وضعت لتعزز ولتطبق على تجمع من البيانات النصية على مستوى أعلى. النظام المقترح يحسن خوارزمية العنقدة المضبب من خلال اقتراح ستراتيجية جديدة لاختيار مراكز التجميع الاولية. لحل مشكلة أختيار مراكز التجميع الاولية والتي تعاني منها خوارزمية العنقدة المضببة القديمة.

عنقدة الوثائق النصية تعني تجميع الوثائق والنصوص المتشابهة الى مجاميع وهذا التجميع للوثائق غير خاضع للرقابة، عند استخراج البيانات المهمة من النص وأستخراج خصائص مهمة. النموذج الأكثر شيوعا لتمثيل الوثائق هو نموذج متجه الفضاء (VSM) الذي يجسد مجموعة من الكلمات المهمة الموجودة في الوثائق.

النظام المقترح يحسن خوارزمية العنقدة المضببة والبيانات بحيث مشكلة الضبابية تم التعبير عنها لبناء المجموعات المضببة المطلوبة، من خلال توظيف خوارزمية العنقدة المضببة وفي مرحلتين، المرحلة الأولى هي البحث عن مركز التجمع، والمرحلة الثانية هي ايجاد الوثائق التابعة لمجموعات البيانات في المجاميع. تم استخدم رويترز 21578 لأغراض تجريبية. النتائج تم مقارنتها والتي تم الحصول عليها بالتركيز على الوقت المستغرق لتجميع البيانات. تم الحصول على نتائج جيدة بالمقارنة مع خوارزمية العنقدة المضببة القديمة وبوقت أقل لتجميع بيانات كبيرة.

**الكلمات المفتاحية:** العنقدة المضببة (الضبابية), مجاميع الوثائق, استخراج المعلومات, رويترز 21578.

**Abstract**

The enhance FCM is place advancing and useful to agreement through text document, clustering happening highest of the outmoded FCM. The suggested system progresses the standard fuzzy c-means algorithm (FCM) through adopting a new approach aimed at choosing the preliminary cluster centers, to resolve the tricky that the outdated fuzzy c-means (FCM) clustering algorithm has effort in choosing the preliminary cluster centers.

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 system in this paper improves fuzzy problem by construct required fuzzy algorithm and applied it in text document clustering, the subjugated through the fuzzy clustering algorithm as dual phase, the chief phase is the principal command instant (i.e. prototype of corpus) and the additional phase is to describe the limitations of the documents of the clusters. Reuters 21578 dataset is cast-off for untried purpose. The effect was acquired by matched with the time essential to complete clustering algorithm. The suggested system originate actually a worthy outcome as matched with the others by consuming time that essential to complete clustering huge datasets similar Reuters 21578.

**Keywords:**Fuzzy clustering, documents datasets, information extraction, Reuters 21578.

**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], 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]. The clustering document was development of mechanically assemblage the connected papers into clusters. As an alternative of thorough complete documents for applicable information relevant document can be capably regained and retrieved by earnings of document clustering [3]

Between the FCM approaches [9] is the greatest healthy-identified technique for it needs the benefit of strength for uncertainty and preserves abundant additional material than rather solid clustering approaches. The procedure was an allowance of traditional besides the brittle k-means algoritm in fuzzy set area. It is extensively intentional and functional in design appreciation, image separation and data mining and so on [4, 5].

Fuzzy C-Means was used as algorithm aim at clustering the documents, which modify the prototypes of the clusters definite originally, the divider matrix

generous the association degree of individually document to each cluster. To decrease the dissimilarity among a document and a cluster prototype was tried by update [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, Vilas V. Pichad, and Sachin N Deshmukh, [15], presented the current digital word technologies information in computer world. In this document clustering algorithms for digital forensic analysis of computers seized devices play important role in real word investigation case conducted by police investigations. Document clustering for forensic analysis is used to study the source and content of various messages as evidence, the results of an experimental which compare the approaches using document clustering algorithm. In the Digital forensic analysis of investigation, there are total three well-known document clustering algorithms are used. It has been implemented to be used with k-means, Average link and complete link algorithm. Experimental result show that formulation can perform very well on large data to be clustered in compute forensic to overcome this problem.
- 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 aimed at applicable material, these clusters resolve advance the competence and evade touching of innards. Birch hierarchical clustering algorithm that container stay useful to at all social clustering difficult, and its request to some non-decision corpus has exposed his presentation toward remain similar to k-means standards permit designs to fit to altogether the

clusters through opposing grades of association. The authors compared their effort through firm clustering using birch algorithms. The result of comparing shows that their effort evades happy intersection and capable to attain greater routine to k-means once superficially appraised on an inspiring dataset of legendary passages. In their proposed system, they used birch clustering algorithm that functions on cluster twitch with original verge and supplements 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 excavated document is pre-processed was presented by the authors, now the document was ordered with frequency of apiece word that can be designed by via term frequency-inverse document frequency method. After that the alike info is gathered consuming fuzzy c-mean clustering which has been experimentally showed and confirmed through the outcomes. Their proposed enhanced the clustering accurateness and it was fewer cataloguing 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 abstraction since diverse web pages was deliberate which was completed lacking slightly human data effort. A pattern was healthy clear which would suggest the outline to be secondhand to label just how the standards were introduced into the pages. The removal algorithm is to excerpt values as of web pages. This algorithm was skilled to make the pattern denoting clear set of words as shared incidence. The authors tool was MDL attitude to achieve the unidentified amount of clusters then MinHash method has been applied to haste the clustering procedure. The new outcomes display that their effort was effectually cluster web documents.
- In 2013, M.Y.Gong, W.Ma Liang [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 contract through meteorological data on upper of the traditional fuzzy c mean. The authors presents the structures and the mining procedure of the exposed basis data mining stage WEKA. The experimental results show that the proposal was generated improved clustering outcomes than k-means algorithm and the outmoded 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 job of assemblage a regular of items in such a method that items in the similar collection are additional alike to apiece additional than to those in additional collections. The authors shows that their suggested document clustering technique is practically useful document clustering method through great intra-match and little inter-match universe consuming a fleeting review on optimization process to text document clustering usage of exterior cause similar 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 suggested system practices the methods for document clustering to enable the huge corpus predictors to organize their effort powerfully.

The proposed method holds of twin phases: training phase and testing phase. The training phase impartial is to alter charge of selected model vectors affording to a usual of documents $d_i = \{d_1, d_2, \ldots, dn\}$ which is training papers, all document consistent its feature vectors (set of terms in each document around 5509 terms (features)), whereas the objective of the testing phase is to cluster the received documents into constraint clusters based on the center vectors fashioned since the training phase. Figure (1) shows the block diagram of the suggested method.

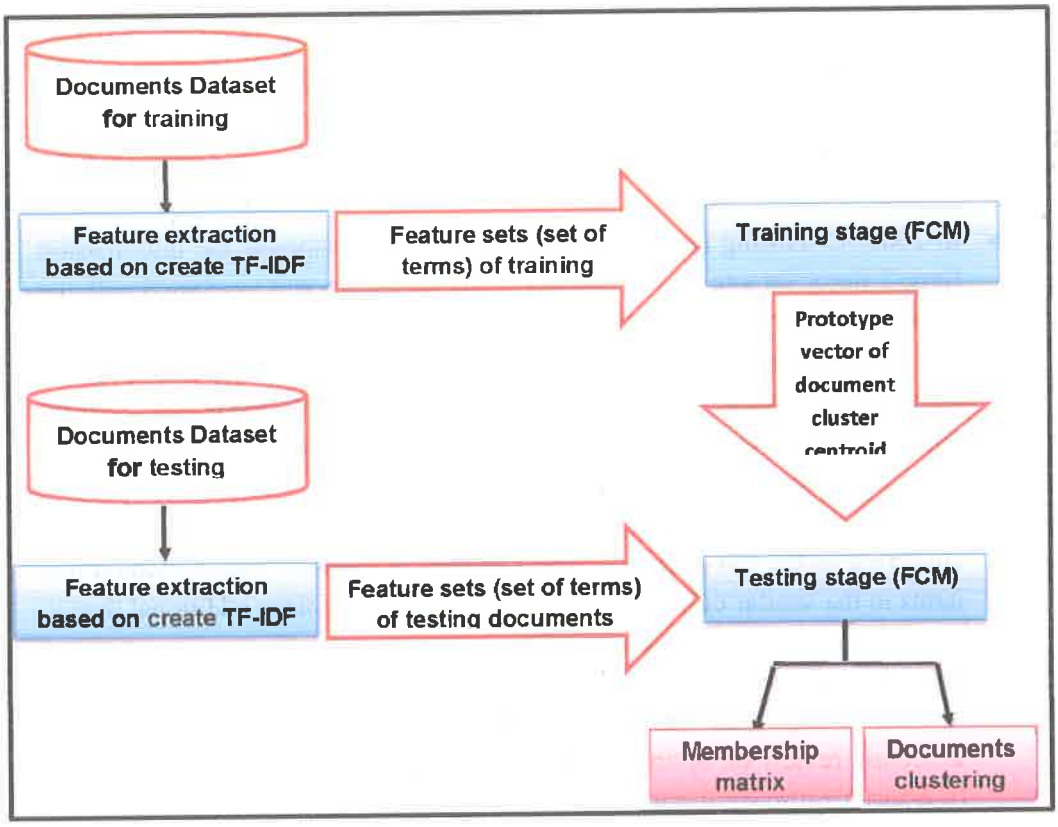


Figure1: The block diagram of the proposed method

### 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**

Whole documents was collect 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 vector space model (VSM). Assumed di was a document stayed restrained as direction in VSM, di, the list of terms as exposed, in its diffident process, separately document is personified via the TF vector, $d_{tf} = \{tf_1, tf_2, \dots\dots, tf_n\}$, Where $tf_i$ is the amount term i in the document. As follows:

$$\text{Terms Frequency (TF)} = \frac{\text{Number of times terms T appears in a document}}{\text{total number of terms in the document}} \quad \dots\dots (1)$$

and the Inverse Document Frequency ( IDF)

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

For each term in document the proposed system calculate TF-IDF value.

After extract features, the VSM matrix was clustered used the improvement FCM.

**3.3. Clustering with improvement fuzzy C mean algorithm**

The proposed method was used a technique to define document prototype that is based on firstly randomly select for prototype document, 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 thus, 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:

**Training phase algorithm**

**Input:**

- Datasets to stay clustering.
- Clusters number.
- Parameter of fuzziness degree.
- Threshold $To_1 > To_2$
- IT=1 as number of iteration.
- Iteration, maxi as the greatest iteration.

**Output:**

- Center vectors for each clusters ($C_i$).

**Procedures begin:**

Step 1: Extract the document datasets, split document in to n category corresponded to user supervision at the selecting features based on the tags that appear in documents, Then preprocessing the documents content, tokenization the document, eliminate the stop words and annoying words, stemming the words and kept the processed documents as $D_i$, where i= 1,2,3.... N.

Step 2: Construction of document-term matrix and discovery TF-IDF matrix of $D_i$, anywhere 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

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

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

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

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 documentsprototypes
   $c_1, c_2, \ldots\ldots c_i$ , where i = number of required cluster.
f) cluster membership values was calculated as :

$$U_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{\|d_k - c_i\|}{\|d_k - c_j\|}\right)^{\frac{2}{m-1}}} \quad \ldots\ldots (3)$$

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

Where $w_{ik}$ is TF-IDF value of document k vector, $\|d_k - c_i\|$ reresent the Euclidean distance between document $_k$ and the document center.

and

$$\|d_k - c_j\| = \left\| \begin{bmatrix} w_{1k} \\ w_{2k} \\ w_{3k} \\ \dots \dots \dots \\ w_{Nk} \end{bmatrix} - \begin{bmatrix} c_{1j} \\ c_{2j} \\ c_{3j} \\ \dots \dots \\ c_{Nj} \end{bmatrix} \right\| \dots\dots\dots (5)$$

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

Which embody the euclidean distance among document $_k$, per every document center vector j where j = {1,2, …. , number of document clusters}.

**Step 5**: modify document center of the essential clusters by Eq. 6:

$$C_j = \frac{\sum_{i=1}^{n} [U_{ij}]^m * d_i}{\sum_{i=1}^{n} [U_{ij}]^m} \dots\dots\dots (6)$$

j was the number of clusters from 1to c.

i was the number of document from 1 to n.

$U_{ij}$ was the membership degree of document i in the cluster j.

$$\mathbf{C_j} = \frac{U_{1j}{}^m * w_{i1} + U_{2j}{}^m * w_{i2} + \dots\dots\dots + U_{nj}{}^m * w_{ij}}{U_{1j}{}^m + U_{2j}{}^m + U_{3j}{}^m + \dots\dots\dots + U_{nj}{}^m} \dots\dots (7)$$

**Step 6:** The stopping criteria was checked, by iteration number greater then $max_i$ then stop, else increase iteration number, and go to step 3.

End.

Main steps of testing phase is shown in algorithm (2):

**Algorithm 2: Testing stage**

**Input:**

- Testing dataset (DT2).
- The amount of clusters.
- The parameter of fuzziness.
- Center vectors starting from training stage.
- The number of iteration was set, IT=1.

**Output:**

- Clustering set.
- Matrix of membership degree.

**Procedures begin:**

**Step 1**:

- The document center $C_1, C_2, \ldots, Ci$ from training phase and all input document $d_1, d_2, \ldots, d_k$, calculate cluster membership $U_{ik}$ as :

$$U_{ik} = \frac{1}{\sum_{j=1}^{c} \left( \frac{\|d_k - c_i\|}{\|d_k - c_j\|} \right)^{\frac{2}{m-1}}} \quad \ldots\ldots\ldots (8)$$

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

which embody the euclidean distance among document k, and the document center vector i. And

$$\|d_k - c_j\| = \left\| \begin{bmatrix} w_{1k} \\ w_{2k} \\ w_{3k} \\ \ldots \ldots \ldots \\ w_{Nk} \end{bmatrix} - \begin{bmatrix} c_{1j} \\ c_{2j} \\ c_{3j} \\ \ldots \ldots \\ c_{Nj} \end{bmatrix} \right\| \ldots\ldots\ldots (10)$$

***where*** **j** was amount of cluster from 1,2,.... C , $\|d_k - c_j\|$ was the euclidean distance among document k, and all document center vector j , the j was number of document clusters

**Step 2**: modify centers using:

$$C_j = \frac{\sum_{i=1}^{n} [U_{ij}]^m * d_i}{\sum_{i=1}^{n} [U_{ij}]^m} \ldots\ldots\ldots (11)$$

j was the number of clusters

i was the number of document

$U_{ij}$ was the membership degree of document i in the cluster j.

$$C_j = \frac{U_{1j}^{m}[w_{i1}] + U_{2j}^{m}[w_{i2}] + \cdots\cdots + U_{nj}^{m}|w_{ij}|}{U_{1j}^{m} + U_{2j}^{m} + U_{3j}^{m} + \cdots\cdots + U_{nj}^{m}} \ldots\ldots\ldots (12)$$

**Step 3:** the tag $Cluster_1$, $Cluster_2$, ……, $Cluster_j$ was assign to the tested document di, i= 1,2, …. n.

$$Dj = \begin{cases} cluster_1 & if\ U_{1j} > U_{\ldots} \\ cluster_2 & if\ U_{2j} > L \\ \ldots\ldots\ldots & \\ cluster_n & if\ U_{nj} > otherwise \end{cases} \quad \ldots\ldots\ldots\ldots (13)$$

End.

**4. Implementation FCM algorithm**

The detail information of the proposed system experiments was described in the followed detail example. The improved FCM algorithm was implemented on the platform Reuters 21578 datasets. The proposed system experimental data shows in the table (1):

*Table 1: Experimental Datasets Summarize*

| Dataset | **Object (document)** | **Dimensions** | **Missing** |
|---|---|---|---|
| **Reuters 21578** | 16000 | 3 | Null |

Simple Example of membership matrix calculation

1- We have initial centroid $\begin{matrix} 3 & 11 \\ 2 & 2 \\ 4 & 3 \end{matrix}$ (with m=2). From the simple TF-IDF matrix =

$$\begin{bmatrix} 2 & 3 & 4 & 5 & 11 \\ 3 & 2 & 1 & 6 & 2 \\ 1 & 4 & 4 & 2 & 3 \end{bmatrix}$$

2- **For first document vector** $\begin{pmatrix}2\\3\\1\end{pmatrix}$ **:**

$$U_{ki} = \frac{1}{\Sigma_{j=1}^{c}\left(\frac{\|d_k - v_i\|}{\|d_k - v_j\|}\right)^{\frac{2}{m-1}}}$$

$$U_{d1}C_1 = \frac{1}{\left[\frac{\left\|\begin{matrix}2&3\\3&-2\\1&4\end{matrix}\right\|}{\left\|\begin{matrix}2&3\\3&-2\\1&4\end{matrix}\right\|} + \frac{\left\|\begin{matrix}2&3\\3&-2\\1&4\end{matrix}\right\|}{\left\|\begin{matrix}2&11\\3&-2\\1&3\end{matrix}\right\|}\right]^{\frac{2}{2-1}}}$$

$\left\|\begin{matrix}2&3\\3&-2\\1&4\end{matrix}\right\|$ = is the distance equal to square root of Euclidean distance Between first document and cluster center 1.

$= (2-3)^2 + (3-2)^2 + (1-4)^2 = 11$

The distance regarding to cluter1 $= \sqrt{euclidean\ distance}$

The distance regarding to cluter1 $== \sqrt{11} = 3.316$

$\left\|\begin{matrix}2&11\\3&-2\\1&3\end{matrix}\right\|$ = is the distance equal to square root of Euclidean distance

The first document and cluster center 2.

$= (2-11)^2 + (3-2)^2 + (1-3)^2 = 85$

The distance regarding to cluter2 $== \sqrt{85} = 9.219$

$$U_{d1}C_1 = \frac{1}{\left(\left(\frac{3.316}{3.316}\right) + \left(\frac{3.316}{9.219}\right)\right)^{\frac{2}{2-1}}} = 1.846$$

(The membership of first document to first cluster)

$$U_{d1}C_2 = \frac{1}{\left[\frac{\left\|\begin{matrix}2&11\\3&-2\\1&3\end{matrix}\right\|}{\left\|\begin{matrix}2&3\\3&-2\\1&4\end{matrix}\right\|} + \frac{\left\|\begin{matrix}2&11\\3&-2\\1&3\end{matrix}\right\|}{\left\|\begin{matrix}2&11\\3&-2\\1&3\end{matrix}\right\|}\right]^{\frac{2}{2-1}}}$$

$\left\|\begin{matrix}2&3\\3&-2\\1&4\end{matrix}\right\|$

= is the distance equal to square root of Euclidean distance Between first document and cluster center 1.

$= (2-3)^2 + (3-2)^2 + (1-4)^2 = 11$

The distance regarding to cluter1 = $\sqrt{euclidean\ distance}$

The distance regarding to cluter1 == $\sqrt{11}$ =3.316

$\left\| \begin{matrix} 2 \\ 3 \\ 1 \end{matrix} - \begin{matrix} 11 \\ 2 \\ 3 \end{matrix} \right\|$ = is the distance equal to square root of Euclidean distance Between the first document and cluster center 2.

$= (2-11)^2 + (3-2)^2 + (1-3)^2 = 85$

The distance regarding to cluter2 == $\sqrt{85}$ = 9.219

$$U_{d1}C_2 = \frac{1}{\left(\left(\frac{9.316}{3.316}\right)+\left(\frac{9.219}{9.219}\right)\right)^{\frac{2}{2-1}}} = 14.5$$

(The membership of first document to second cluster)

*Table 3: membership matrix of first document*

| | Document1 | Document2 | ..... | Document N |
|---|---|---|---|---|
| *Cluster center 1* | **1.846** | | ..... | |
| *Cluster center 2* | **14.5** | | ..... | |

- **For document$_2$** $\begin{pmatrix} 3 \\ 2 \\ 4 \end{pmatrix}$ **(2nd document ):**

**$U_{d2}C1$** = 100%, the membership of second document to first cluster

**$U_{d2}C2$**= 0%, the membership of second document to second cluster

*Table 4: membership matrix of first document and second document*

| | Document1 | Document2 | ..... | Document N |
|---|---|---|---|---|
| *Cluster center 1* | **1.846** | **100%** | ..... | |

*Cluster center 2* | **14.5** **0%** .....

**For document$_3$** $\begin{pmatrix}4\\1\\4\end{pmatrix}$ **(3rd document):**

$$U_{d3}C_1 = \frac{1}{\left[\frac{\left\|\begin{matrix}4&3\\1&-2\\4&4\end{matrix}\right\|}{\left\|\begin{matrix}4&3\\1&-2\\4&4\end{matrix}\right\|} + \frac{\left\|\begin{matrix}4&3\\1&-2\\4&4\end{matrix}\right\|}{\left\|\begin{matrix}4&11\\1&-2\\4&3\end{matrix}\right\|}\right]^{\frac{2}{2-1}}}$$

$\left\|\begin{matrix}4&3\\1&-2\\4&4\end{matrix}\right\|$ = is the distance equal to square root of Euclidean distance Between third document and cluster center 1.

$= (4-3)^2 + (1-2)^2 + (4-4)^2 = 4$

The distance regarding to cluter1 = $\sqrt{euclidean\ distance}$

The distance regarding to cluter1 == $\sqrt{4} = 2$

$\left\|\begin{matrix}4&11\\1&-2\\4&3\end{matrix}\right\|$ = **is** the distance equal to square root of Euclidean distance Between the third document and cluster center 2.

$= (4-11)^2 + (1-2)^2 + (4-3)^2 = 51$

The distance regarding to cluter2 == $\sqrt{51} = 7.141$

$$U_{d3}C_1 = \frac{1}{\left(\left(\frac{2}{2}\right)+\left(\frac{2}{7.141}\right)\right)^{\frac{2}{2-1}}} = 0.610$$

(The membership of third document to first cluster)

$$U_{d3}C_2 = \frac{1}{\left[\frac{\left\|\begin{matrix}4&11\\1&-2\\4&3\end{matrix}\right\|}{\left\|\begin{matrix}4&3\\1&-2\\4&4\end{matrix}\right\|} + \frac{\left\|\begin{matrix}4&11\\1&-2\\4&3\end{matrix}\right\|}{\left\|\begin{matrix}4&11\\1&-2\\4&3\end{matrix}\right\|}\right]^{\frac{2}{2-1}}}$$

$\left\|\begin{matrix}4&3\\1&-2\\4&4\end{matrix}\right\|$ = is the distance equal to square root of Euclidean distance Between third document and cluster center 1.

$= (4-3)^2 + (1-2)^2 + (4-4)^2 = 4$

The distance regarding to cluter1 == $\sqrt{4} = 2$

$\left\| \begin{matrix} 4 & 11 \\ 1 & 2 \\ 4 & 3 \end{matrix} \right\|$ = **is** the distance equal to square root of Euclidean distance

Between the third document and cluster center 2.

$= (4-11)^2 + (1-2)^2 + (4-3)^2 = 51$

The distance regarding to cluter2 == $\sqrt{51} = 7.141$

$$U_{d3}C_2 = \frac{1}{\left(\left(\frac{7.141}{2}\right)+\left(\frac{7.141}{7.141}\right)\right)^{\frac{2}{2-1}}} = 0.047$$

(The membership of third document to second cluster)

*Table 4: membership matrix of first document ,second document, and third document*

| | Document1 | Document2 | Document3 | Document 4 |
|---|---|---|---|---|
| *Cluster center 1* | **1.846** | **100%** | **0.610** | |
| *Cluster center 2* | **14.5** | **0%** | **0.047** | |

- **For document4** $\begin{pmatrix} 5 \\ 6 \\ 2 \end{pmatrix}$ **(4th document):**

$$U_{d4}C1 = \frac{1}{\left[\frac{\left\| \begin{matrix} 5 & 3 \\ 6 & 2 \\ 2 & 4 \end{matrix} \right\|}{\left\| \begin{matrix} 5 & 3 \\ 6 & 2 \\ 2 & 4 \end{matrix} \right\|} + \frac{\left\| \begin{matrix} 5 & 3 \\ 6 & 2 \\ 2 & 4 \end{matrix} \right\|}{\left\| \begin{matrix} 5 & 11 \\ 6 & 2 \\ 2 & 3 \end{matrix} \right\|}\right]^{\frac{2}{2-1}}} =$$

$\left\| \begin{matrix} 5 & 3 \\ 6 & 2 \\ 2 & 4 \end{matrix} \right\|$ = is the distance equal to square root of Euclidean distance

Between fourth document and cluster center 1.

$= (5-3)^2 + (6-2)^2 + (2-4)^2 = 24$

The distance regarding to cluter1 == $\sqrt{24} = 4.89$

$\left\| \begin{matrix} 5 & 11 \\ 6 & 2 \\ 2 & 3 \end{matrix} \right\|$

= **is** the distance equal to square root of Euclidean distance

Between the fourth document and cluster center 2.

$$= (5-11)^2 + (6-2)^2 + (2-3)^2 = 53$$

The distance regarding to cluter2 == $\sqrt{53} = 7.28$

$$U_{d4}C_1 = \frac{1}{\left(\left(\frac{4.89}{4.89}\right)+\left(\frac{4.89}{7.28}\right)\right)^{\frac{2}{2-1}}} = 0.35$$

(The membership of fourth document to first cluster)

$$U_{d4}C2 = \frac{1}{\left[\frac{\left\|\begin{matrix}5 & 11\\ 6-2\\ 2 & 3\end{matrix}\right\|}{\left\|\begin{matrix}5 & 3\\ 6-2\\ 2 & 4\end{matrix}\right\|} + \frac{\left\|\begin{matrix}5 & 11\\ 6-2\\ 2 & 3\end{matrix}\right\|}{\left\|\begin{matrix}5 & 11\\ 6-2\\ 2 & 3\end{matrix}\right\|}\right]^{\frac{2}{2-1}}} =$$

$$\left\|\begin{matrix}5 & 11\\ 6-2\\ 2 & 3\end{matrix}\right\| = (5-11)^2 + (6-2)^2 + (2-3)^2 = 53$$

The distance regarding to cluter2 == $\sqrt{53} = 7.28$

$$U_{d4}C_2 = \frac{1}{\left(\left(\frac{7.28}{4.89}\right)+\left(\frac{7.28}{7.28}\right)\right)^{\frac{2}{2-1}}} = 0.16$$

(The membership of fourth document to second cluster)

*Table 5:- The final membership of each document*

| | Document1 | Document2 | Document3 | Document 4 |
|---|---|---|---|---|
| *Cluster center 1* | **1.846** | **100%** | **0.610** | **0.35** |
| *Cluster center 2* | **14.5** | **0%** | **0.047** | **0.16** |

*Table 6: Final Clusters from FCM*

| *Cluster 1* | Cluster 2 |
|---|---|

| | |
|---|---|
| *Document2* | Document1 |
| *Document3* | |
| *Document4* | |

**5. 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.

**5.1. 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**

Was a quantity for the degree of each cluster holds single class tag. To calculate purity, for cluster $j$, the amount of happenings for separately class $i$ were calculated and choice the determined amount $max_{ij}$, the purity was the summary of totally greatest appearance $max_{ij}$ divided by the whole amount of documents $n$.

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

- **Entropy**

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

$$\mathbf{E(j)} = \sum_{i=1}^{c} P_{ij} \log_2 \frac{1}{P_{ij}} \dots\dots\dots (14)$$

where, $c$ is the quantity of classes, $P_{ij}$ is the likelihood that associate of cluster $j$ be appropriate to class i,

$$\mathbf{P_{ij}} = \frac{\mathbf{n_{ij}}}{\mathbf{n_j}} \dots\dots\dots (15)$$

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:

$$\mathbf{E} = \sum_{\mathbf{j=1}}^{\mathbf{k}} \frac{\mathbf{n_j\, E(j)}}{\mathbf{n}} \dots\dots\dots (16)$$

where $k$ is the quantity of clusters, the total number of objects in cluster $j$ is $n_j$ , 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 $j$.

$$\mathbf{Recall(i,j)} = \frac{n_{ij}}{n_i} \quad \ldots\ldots\ldots\ldots (17)$$

$$\mathbf{accuracy\ (I,j)} = \frac{n_{ij}}{n_j} \quad \ldots\ldots\ldots\ldots (18)$$

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

$$\mathbf{F(i,j)} = \frac{(2*\mathbf{Recall(I,J)}*\mathbf{accuracy\ (I,J)})}{\mathbf{Recall(I,J)}+\mathbf{accuracy(I,J)}} \quad \ldots\ldots\ldots\ldots (19)$$

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

$$F = \sum_{i=0}^{c} \frac{n_i}{n} \text{ Max } F(i,j) \quad \ldots.. (20)$$

**5.2 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:

$$\text{Maximize} \frac{\sum_{j=1}^{k}\sum_{d=0}^{n_j}\cos(d_j,\ \cdot\ }{n} \quad \ldots\ldots\ldots\ldots (21)$$

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.

**6. The Results of experimental**

The Reuters 21578 datasets was used for fuzzy clustering tests about 1000 selected documents for clustering them, actual number of classes 40. Table 7 shows the setting for the suggested system experiment.

**Table 7: Setting for Experiment**

| | | |
|---|---|---|
| Fuzzy C Mean parameters | The cluster number | Was set randomly |
| | The fuzzier parameters | Was set randomly |
| | The used distance | Was euclidean distance |
| | The membership weights | Was set randomly |
| | Stopping criteria | Stopping criteria < 0.005 |

Table 8 shows how the proposed system perform the external measures with using improvement FCM.

**Table 8: 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 |
|---|---|---|---|---|---|
| **Improvement FCM algorithm** | 0.79 | 0.58 | 0.46 | 0.6 | 0.75 |
| **Traditional FCM algorithm** | 0.74 | 0.52 | 0.40 | 0.50 | 0.71 |
| **Entropy** | **C=2** | **C=3** | **C=4** | **C=5** | **C=6** |
| **Improvement FCM algorithm** | 1.50 | 1.13 | 0.90 | 1.13 | 1.22 |
| **Traditional FCM algorithm** | 1.56 | 1.15 | 0.95 | 1.2 | 1.5 |
| **F-measures** | **C=2** | **C=3** | **C=4** | **C=5** | **C=6** |
| **Improvement FCM algorithm** | 1.59 | 1.68 | 1.79 | 1.88 | 1.91 |
| **Traditional FCM algorithm** | 1.54 | 1.62 | 1.70 | 1.83 | 1.82 |

Table 9 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) using improvement FCM algorithm, designed for apiece rate of C present is an best value of α giving the best clustering quality.

**Table 9: The Purity measures with varied C and α on subset TD1 for improvement FCM algorithm.**

| 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 |

## 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 different technique to regulate cluster prototype, then it was generous a 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 advance the presentation of the FCM in the arena of another datasets since additional aspects.

## References

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

[2] C. C. Hung, S. Kulkarni and B. C. kuo, J.selected Topics in signal processing, vol.5, no.3, 2011.

[3] C. Hwang and F. C. H. Rhee, J. "Transactions on Fuzzy Systems", 2007.

[4] 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.

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

**[6]** Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kauffman Publishers, 2014.

[7] M.Gong, Y.Liang, W.Ma and J.Ma, J.IEEE Transaction on Image processing, vol.22, no. 2, 2013.
[8] M.Gong.Y. Liang,W.Ma ,"Transactions on image processing", 2013.
[9] P. Chiranjeevi, T. Supraja, P. Srinivasa Rao, "A Survey on Extension Techniques for Text Document Clustering", 2015.
[10] Pengtao Xie, and Eric P.Xing, "Integrating Document Clustering and Topic Modeling", Tsinghua University, China, 2014.
[11] Rashmi D thakare and Manisha R patil, "Extraction of Template using Clustering from Heterogeneous Web Documents", International Journal of Computer Applications 2015.
[12] Stuti Karol, Veenu Mangat, "Evaluation of text document clustering approach based on particle swarm optimization", 2013.
[13] Tanagra Data Mining Ricco Rakotomalala, "Text mining" with Knime and RapidMiner. Reuters Text Categorization, 2016.
[14] Vilas V Pichad, and 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
[15] Yinghua Lu, Tinghuai Ma and Changhong Yin, "Implementation of fuzzy c mean clustering algorithm in meteorological data", 2013.