An Improved Clustering Method for Text Documents Using Neutrosophic Logic

Nadeem Akhtar, Mohammad Naved Qureshi and Mohd Vasim Ahamad

1 Introduction

As a technique of Information Retrieval, we can consider clustering as an unsupervised learning problem in which we provide a structure to unlabeled and unknown data [1, 2]. Clusters formed as part of clustering contains the objects which are similar to each other in terms of their content [3, 4]. The following example as shown in Fig. 1 clearly depicts how clusters can be formed.

First, we will discuss basics of Fuzzy C Means clustering and then our approaches to modify it to get better results in terms of its accuracy. Fuzzy C Means (FCM) clustering method assigns fuzzy membership for documents belonging to clusters [3, 5]. The fuzzy membership values range between 0 and 1. Therefore, each cluster is considered as the fuzzy set of all documents. It was developed by Dunn in 1973. The Fuzzy C Means clustering method starts by assuming C as the number of clusters required, selecting random cluster centers, and assigning truth membership values to each document with respect to every cluster center. The membership values for each cluster and each document must be equal to one. In each iteration, cluster centers are updated. This algorithm iterates up to minimum objective function which can be define as [6, 7]:

N. Akhtar
Department of Computer Engineering, ZHCET, Aligarh Muslim University, Aligarh 202002, India
e-mail: nadeemakhtar@zhcet.ac.in

M. N. Qureshi
University Polytechnic, Aligarh Muslim University, Aligarh 202002, India
e-mail: navedmohd786@gmail.com

M. V. Ahamad (✉)
Womens Polytechnic, Aligarh Muslim University, Aligarh 202002, India
e-mail: vasim.iu@gmail.com

R. Ali and M. M. S. Beg (eds.), *Applications of Soft Computing for the Web*,
https://doi.org/10.1007/978-981-10-7098-3_10

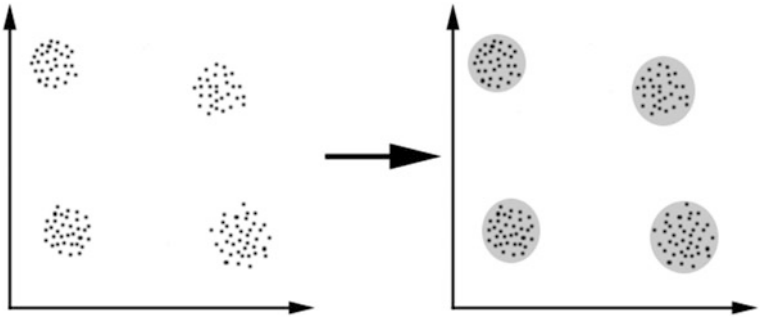

Fig. 1 Clusters of similar objects

$$J_m = \sum_{i=1}^{N} \left(\sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2 \right) \tag{1}$$

where $1 \leq m < \infty$ and $m > 1$, u_{ij} is the degree of membership of ith document x_i with respect to the jth cluster c_j.

The drawback with Fuzzy C Means is that if the degree of membership for a particular document for a cluster is somewhat equal for two clusters so there is ambiguity over here. So here it is difficult to tell how much it is true that document d belongs to a cluster, any cluster x. So to handle this ambiguity we need another term called as indeterminacy value which is provided by Neutrosophic logic. In case of Neutrososphic logic, we have truth, falsity, and indeterminacy values for a single document belonging to a cluster. So based on these three values we can accurately classify the document to a particular cluster, i.e., the document will belong to the cluster when it has high t, i, f value for that cluster as compared to other.

2 Background

2.1 Fuzzy Logic

Fuzzy logic [8] is the expansion of the classical and multivalued logics. It is based on the basic probability theorem that a particular event can have a probability range from 0 to 1. Fuzzy logic allows variables to have values between 0 and 1. The variable is considered to be false if its value is 0 and considered as true for value equals to 1. Fuzzy logic also considers intermediate values such as $x = 0.87$, to incorporate partial truthiness or falsity of a variable. In comparison to classical logic where we have only two outcomes either true or false, in fuzzy logic we can have

various values in between to completely true and false values to deal with partial and uncertain data.

2.2 Neutrosophic Logic

The fuzzy logic is proposed to deal with the vagueness and uncertainty [8, 9]. It has two values associated with each variable, the degree of truthfulness and degree of falsehood. It can be represented as FS = {T, F}, where T and F are the degree of truthiness and degree of falsehood of the variable toward the set FS, respectively. Neutrosophic logic introduces a new parameter to a fuzzy set, called as indeterminacy. Neutrosophic logic theory considers every possible outcome for a variable X like X, Anti-X, and Neut-X which is neither X nor Anti-X [10]. According to this theory if there is indeterminacy for a particular variable or idea than that also can be expressed with a degree of membership for a variable.

3 Proposed Work

In this section, we are introducing document clustering using Neutrosophic logic. This section explores two approaches of data clustering with the help of Neutrosophic logic. The results are very promising and show the possibility of quality improvement in data clustering.

In the first approach, we added the indeterminacy factor of Neutrosophic logic to Fuzzy C Means clustering method and modified the formula which calculates the cluster centers and the truth membership of documents toward clusters. The aim of this approach is to introduce the indeterminacy factor to the Fuzzy C Means Clustering Algorithm and grouping documents (say N) in an input dataset into C clusters. The indeterminacy of documents mainly affected by the clusters which are most similar to that document. Using this concept, we calculated the indeterminacy factor using the average value of closest and second closest membership values of document with corresponding clusters. The modified algorithm tries to associate the document with the cluster having higher truth membership grade and lowest indeterminacy values toward the cluster.

As in traditional Fuzzy C Means clustering, this modified also starts with calculating the cluster centers first. Following is the modified formula for calculating the cluster center:

$$c_j = \frac{\sum_{i=1}^{n} \left(I_{ij} \cdot u_{ij}\right)^m \cdot x_i}{\sum_{i=1}^{n} \left(I_{ij} \cdot u_{ij}\right)^m} \tag{2}$$

where $[u_{ij}]$ is membership value matrix of ith document to jth cluster and $[I_{ij}]$ is indeterminacy value matrix of ith document to jth cluster and x_i is the ith document.

Further, values of membership and indeterminacy can be updated in iteration with the following modified formula:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{||x_i - c_j||}{||x_i - c_k||} \right)^{\frac{2}{m-1}}} \tag{3}$$

$$I_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{||x_i - c_j||}{||x_i - \overline{c_{\text{avg}}}||} \right)^{\frac{2}{m-1}}} \tag{4}$$

$$\overline{c_{\text{avg}}} = \frac{c_{pi} + c_{qi}}{2} \tag{5}$$

where p_i and q_i are the clusters with the largest and the second largest membership values for document D, c is the number of clusters required, x_i is ith document, c_j is jth cluster, and m is weighted factor. In this case, we have assumed the value of m as 2.

The proposed algorithm starts by taking the input dataset having D documents and preprocessing it. As in Fuzzy C Means clustering, this algorithm also takes C (number of clusters required) random values as cluster centers. After that, membership values matrix and indeterminacy matrix is initialized. Then, it tries to associate the document with the cluster having higher truth membership grade and lowest indeterminacy values toward the cluster. This algorithm iterates and updates the cluster centers and indeterminacy values using above-mentioned equations. This algorithm repeats until objective function is optimized, which can be define as:

$$J_m = \sum_{i=1}^{n} \sum_{j=1}^{c} \left(u_{ij}\right)^m \cdot \left(||x_i - c_j||\right)^2 \tag{6}$$

where u_{ij} is membership value of ith document to jth cluster, x_i is the ith document, and c_j is the jth cluster.

The second approach consists of three phases which are shown below in Fig. 2.

3.1 Preprocessing and Data Collection

The objective of this phase is to generate dataset for clustering. The format of dataset is according to standard so that if we apply our method on a preprocessed dataset that we can apply Phase II and Phase III directly. Basic steps for phase 1 are listed below:

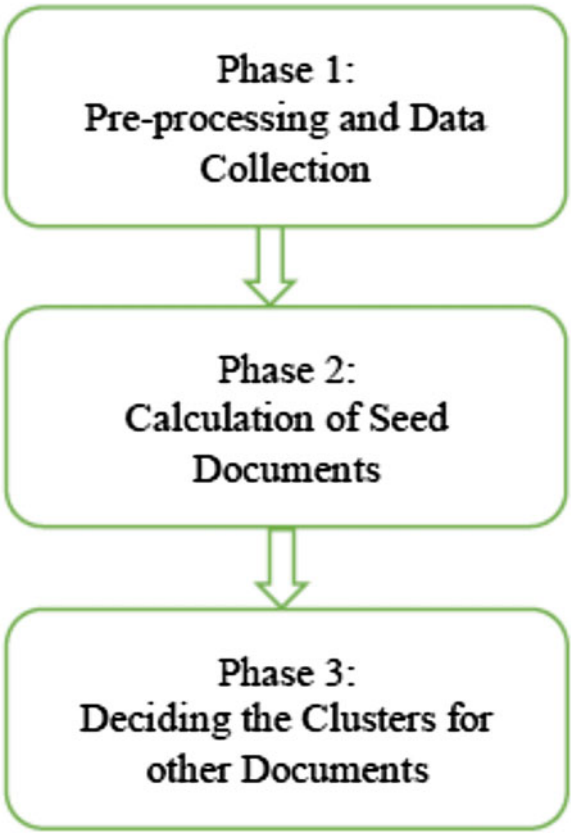

Fig. 2 Different phases of second approach of the proposed algorithm

- Collect URLs of different topics from Google search.
- Get text of URLs
- Remove images and HTML tags from text stream
- Remove helping verbs and stop words
- Perform stemming using Porter's suffix striping algorithm
- Calculate the percentage of appearance of words in a document
- Arrange words and document in dataset.

We perform a Google search using a topic string and then save the URLs of top 100 documents in the search result. We do this for all topics on which we want to generate cluster. Then we extract the data of each url as text string and then remove images and HTML tags. After that we remove helping verbs and stop words. Porter's suffix-stripping algorithm is used for performing stemming over dataset.

The words which are rooting back to the same stem can be considered as same word. For example, "compute", "computing", and "computed" can be stemmed to "comput". After the above step, find out all the words appearing in the document

Table 1 Word dataset

Word dataset	Words id	
Document id	Word id	Frequency
01	1, 2, 4, 5	20, 30, 20, 10
02	2, 4, 5, 6, 3	29, 18, 70, 12
03	5, 3, 2, 4, 14, 9	12, 13, 19, 30, 13, 78

and then calculate the percentage of frequency of words in document. Perform this step for all documents and create a dataset as a table having document id in row side and word id in column side and their respective frequency at the cell position of table as shown in Table 1.

3.2 Calculation of Seed Documents

In this phase, we are deciding the seed for the clusters. These seeds play the role of initial centroid in our algorithm. All other document's cluster is decided with Neutrosophic logic on these seed clusters. The base of these seed documents is Euclidean distance.

Euclidean distance: Euclidean distance can be calculated as the square root of differences between the coordinates of a pair of objects [3, 4]. Each object can be represented as a vector. The Euclidean distance d_{ij} can be calculated using the following formula:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} \left(x_{ik} - x_{jk}\right)^2} \tag{7}$$

where x_{ik} is the kth dimension of ith document, x_{jk} is the kth dimension of jth document, and i and j are n-dimensional vectors.

Basic steps for phase 2 is listed below.

- Select a document randomly and say it as cluster 0
- For $i = 0$ to $n-1$
- Find Euclidean distance (r_i) of all other documents from the cluster i
- Select a document for which $\sum_{k=o}^{i} r_k$ is maximum and say it as cluster $i + 1$
- Consider clusters 1 to n as seeds.

The main motive of this phase is deciding the seed documents from all available documents. Initially, these documents as seed document play a role of the cluster itself. As we can visualize, the probability of being in the same cluster for documents is inversely proportional to the Euclidean distance between two documents. So we are deciding our cluster's seed documents on the basis of the Euclidean distance between documents. We are randomly selecting a document and calling it

as cluster 0. Cluster 0 is a dummy cluster to process seed cluster. As shown in flow diagram, starting from this cluster 0 we find out n more documents as cluster 1 to n. These documents play initial role of clusters. As we are getting more documents in a cluster we are changing the cluster properties.

3.3 Deciding Cluster for Other Documents

After finding out seed documents that act as initial clusters, now the remaining documents have to be assigned to one of the clusters from these seed clusters. We are considering a word as a source of information for deciding the cluster for a document. Definition of some terms to be used in this phase:

- W_{ij} is the percentage of ith word in jth document.
- PC_{ik} is the average of ith word in kth cluster. (Positive Center of ith word in kth cluster).
- NC_{ik} is the average of ith word in all other than kth cluster. (Negative Center of ith word in kth cluster).
- AC_i is the average of ith word in all clusters. (Center of ith word in all clusters).
- R_i = Range of W_i. (maximum (W_i)-minimum(W_i)).

Now we have formulated truth, falsity, and indeterminacy value for a word as used in Neutrosophic logic as defined below:

Truth value for ith word in cluster kth for jth document

$$T_{ijk} = 1 - \frac{(W_{ij} - \mathrm{PC}_{ik})}{R_i} \tag{8}$$

False value for ith word in cluster kth for jth document

$$F_{ijk} = 1 - \frac{(W_{ij} - \mathrm{NC}_{ik})}{R_i} \tag{9}$$

Indeterminate value for ith word for jth document

$$I_{ij} = 1 - \frac{(W_{ij} - \mathrm{AC}_i)}{R_i} \tag{10}$$

Now, these sources of information are not depended on each other so we can combine their T and I, and we can combine their T, I, and F values for a document and cluster. Let us consider the total number of words to be m. Now we are defining some terms.

Truth value for jth document to be in kth cluster is:

$$T_{jk} = 1 - \frac{\sum_{i=1}^{m}\left(\frac{(W_{ij} - \mathrm{PC}_{ik})}{R_i}\right)}{m} \tag{11}$$

False value for jth document to be in kth cluster is:

$$F_{jk} = 1 - \frac{\sum_{i=1}^{m}\left(\frac{(W_{ij} - \mathrm{NC}_{ik})}{R_i}\right)}{m} \tag{12}$$

Indeterminate value for jth document to decide its cluster is:

$$I_j = 1 - \frac{\sum_{i=1}^{m}\left(\frac{(W_{ij} - \mathrm{AC}_{i})}{R_i}\right)}{m} \tag{13}$$

It is clear from these formulas that if a document j is in k^{th} cluster then its corresponding T value should be high and its false value should be low. Therefore, we have introduced a new term "Deciding Factor" as DF given below:

$$\text{Deciding factor (DF)} = T - F \tag{14}$$

But as Truth values are calculated through own documents of a cluster, it should have more weight than False value. So we have modified it as

$$\text{Deciding factor (DF)} = (1.15T) - F \tag{15}$$

DF_{jk} is the "Deciding Factor" of jth document to be in kth cluster

$$\mathrm{DF}_{jk} = \left(1.15T_{jk}\right) - \mathrm{F}_{jk} \tag{16}$$

Algorithm for the phase 3 is given below:

- Calculate the Deciding factor DF for all documents in all clusters.
- Sort all the documents in a cluster according to their DF values for that cluster.
- Select a top 20% of document in all clusters.
- Check any document in these top 20% documents in cluster k is appearing in top 20% documents of any other cluster or not. If yes then set it as claimed (−1) otherwise set it as clear (1).
- Scan all cluster's top 20% list, starting from rank 1, if ith rank documents in all cluster are claimed then check $(i + 1)$th rank.
- Select the clusters whose ith rank document has a clear flag. Set these cluster for these document accordingly. If there is no document in top 20% of all cluster with the clear flag than from first rank documents in all clusters select a cluster whose first rank document has highest DF/I_j value and set it in respective cluster.

- Update the cluster parameters and repeat all the above steps for all other remaining documents.
- End.

Finally, after this algorithm is over, the documents are assigned to their respective clusters with higher truth value and lower indeterminacy with respect to cluster centers. In the following section below we have shown the experimental results of our methodology as compared to Fuzzy C Means clustering using Neutrosophic logic. We have also calculated and compared the accuracy in terms of precision and recall values of both of the algorithm discussed.

4 Result Evaluation

4.1 Dataset Description

We have executed the proposed algorithm on three datasets as listed in Table 2. The dataset 1 is a subset of mini newsgroup dataset available at UCI machine learning database. In this, we have 10 newsgroups having a total of 1000 evenly distributed documents to check the precision variance of the documents. Dataset 2 is also a subset of mini newsgroup dataset but in dataset 2, we put a total of 995 unevenly distributed documents in the 10 newsgroups. The dataset 3 is collected via Google search as described in Sect. 3.1.

4.2 Performance Evaluation

Precision: Peterson and Hearst gave a definition in which they defined precision as a sum of precision of relevant document viewed divide by the total number of documents, viewed, or not viewed. They treated each cluster as a category

Table 2 Description of datasets used

Dataset	#documents	Clusters	Source
Dataset 1	1000	10	https://archive.ics.uci.edu/ml/machine-learning-databases/20newsgroups-mld/
Dataset 2	995	10	https://archive.ics.uci.edu/ml/machine-learning-databases/20newsgroups-mld/
Dataset 3	1000	10	Collected via Google search using 10 different topics on which we wanted to generate cluster

dynamically generated by their method and each category in document cluster is treated as a class. The precision formula for cluster "y" and class "x" is as follows:

$$P(x, y) = \frac{N_{xy}}{N_y} \tag{17}$$

Here N_y is total number of documents in cluster "y" and N_{xy} is total number of common documents in cluster "y" and class "x".

Recall: Recall is defined as the sum of total number of documents common in cluster "y" and class "x", divided by the total number of documents in class "x".

The recall formula for cluster "y" and class "x" is as follows:

$$R(x, y) = \frac{N_{xy}}{N_x} \tag{18}$$

Here N_x is total number of documents in class 'x' and N_{xy} is total number of common documents in cluster "y" and class "x".

F-Measure: F-Measure is a quality measure, having collective impact of Precision and Recall. It combines the Precision and Recall values for each cluster with their corresponding class. The F-Measure formula for cluster "y" and class "x" is as follows:

$$F(x, y) = \frac{2 * P(x, y) * R(x, y)}{P(x, y) + R(x, y)} \tag{19}$$

Here $P(x, y)$ is precision of cluster "y" and class "x" and $R(x, y)$ is Recall of cluster "y" and class "x".

As we can see in Figs. 3, 4, and 5 above, the comparison made between Fuzzy C Means clustering algorithm and both the approaches we discussed. The accuracy of

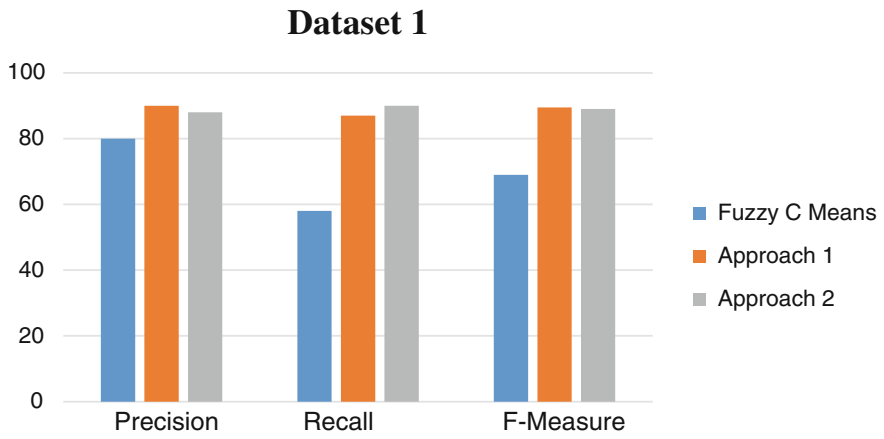

Fig. 3 Accuracy comparison of both approaches with FCM on dataset 1

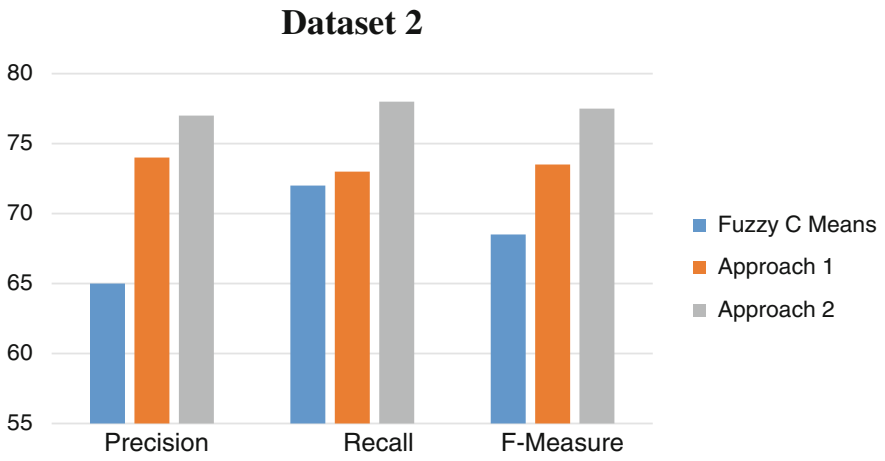

Fig. 4 Accuracy comparison of both approaches with FCM on dataset 2

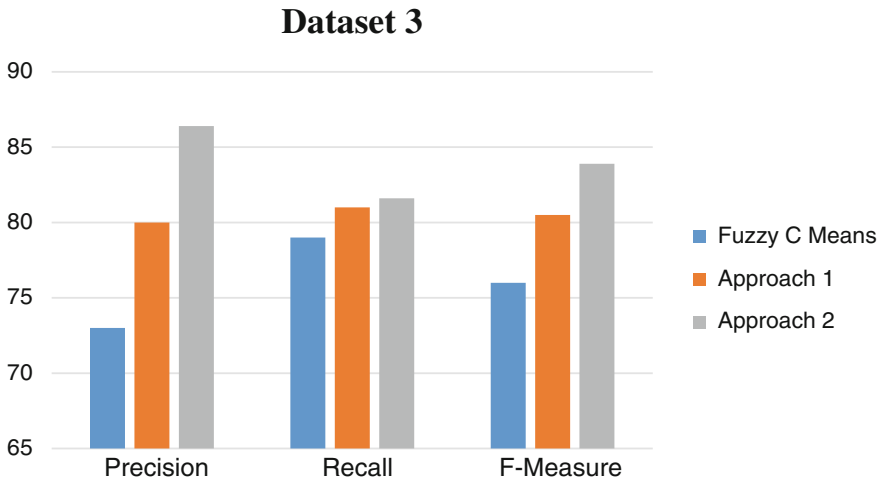

Fig. 5 Accuracy comparison of both approaches with FCM on dataset 3

both the approaches is far better than Fuzzy C Means clustering algorithm in terms of various evaluation parameters (precision, recall, and f-measure). It is clear that modified Fuzzy C Means clustering algorithm is much more accurate than the traditional Fuzzy C Means algorithm for clustering text documents.

5 Conclusion and Discussion

The improved Fuzzy C Means clustering algorithm, we have proposed in this paper, has outperformed the traditional Fuzzy C Means algorithm. With the help of clusters, we can organize text documents which are similar at a single place and it helps us to group other unknown documents in future, to be assigned to one of the known cluster based on the similarity measure. We have proposed two approaches for clustering using Neutrosophic logic. While using fuzzy logic we take into account only two values; degree of truth and degree of falsity, whereas, in Neutrosophic logic, a new factor called as indeterminacy is also involved. Indeterminacy applies to the situation when for a particular document it is not sure that to which cluster it belongs.

In the first approach, we added the indeterminacy factor of Neutrosophic logic to Fuzzy C Means clustering method and modified the formula which calculates the cluster centers and the truth membership of documents toward clusters. The indeterminacy of documents is largely affected by the document clusters that are most similar to that document. Using this concept, we calculated the indeterminacy factor using the average value of closest and second closest membership values of the document with corresponding clusters. The modified algorithm tries to associate the document with the cluster having higher truth membership grade and lowest indeterminacy values toward the cluster.

The second approach has three phases. First, generate the dataset according to the relative frequency of words in a document. Second, decide seed documents for different clusters with the help of Euclidean distance between different documents. Finally calculate the T, I, and F values for all documents with respect to all clusters. Then decide the cluster for each document on the basis of T, I, and F values. In Fuzzy C Means we have to decide the number of clusters prior to clustering but in our methodology, we have found out the seed clusters on the basis of which we can accurately assign a document to a particular seed cluster based on the similarity in contents of that document and the seed cluster. Here we have used Neutrosophic logic where we can model all the three values, i.e., truth, false, and indeterminacy for a document and for a cluster also.

With the help of which we can accurately assign a document to most closed cluster in terms of its contents. We have calculated truth, false, and indeterminate value for every document in a cluster and also for every cluster of documents with the help of which it became very accurate measure for clustering documents in one of the best possible cluster. Also, the introduction of new term, i.e., deciding factor for a document has given more weightage for a document to be in a cluster if its membership value is greatest with respect to others. Finally, the ranking of documents based on the deciding factor has helped to easily cluster them in one of the best possible clusters.

References

1. Hartigan JA (1975) Clustering algorithms. Wiley, London
2. Olson, DL, Delen D (2008) Advanced data mining techniques, 1st edn. Springer, Berlin, p 138. (February 1, 2008), ISBN 3-540-76916-1
3. Akhtar N, Ahamad MV (2015) A modified fuzzy C means clustering using neutrosophic logic. In: Proceedings of IEEE fifth international conference on communication systems and network technologies (CSNT). ISSN/ISBN 978-1-4799-1797-6/15, 10.1109/CSNT.2015.164, pp 1124–1128
4. Hartigan JA, Wong MA (1979) Algorithm AS 136: a *K*-means clustering algorithm. J Royal Stat Soc, Ser C 28(1):100–108. JSTOR 2346830
5. Suganya R, Shanthi R (2012) Fuzzy C-means algorithm—a review. Inter J Sci Res Publ 2(11)
6. Bezdek J (1981) Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York
7. Jain AK, Dubes RC (1988) Algorithms for clustering data. Prentice Hall, Upper Saddle River, NJ. ISBN:0-13-022278-X
8. Zadeh L (1965) Fuzzy sets. Inf Control 8:338–352
9. Dunn J (1973) A fuzzy relative of the Isodata process and its use in detecting compact, well-separated clusters. J Cybern 3(3):32–57
10. Smarandache F (1998) Neutrosophy / neutrosophic probability, set, and logic. American Research Press, Rehoboth, NM
11. Bezdek J, Hathaway R (1988) Recent convergence results for the fuzzy c-means clustering algorithms. J Classif 5(2):237–247