Type2 IFC with SOA for Topic Detection and Document Clustering Analysis

Perumal P (✉ perumalpcse@gmail.com)

Sri Ramakrishna Engineering College

Mathivanan B

Sri Ramakrishna Engineering College

Research Article

Keywords: document clustering, topic detection, feature extraction, term frequency, inverse document frequency, mutual information

Posted Date: August 9th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-743089/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Type2 IFC with SOA for Topic Detection and Document Clustering Analysis

¹*Dr. P. Perumal and ²Dr. B. Mathivanan,

¹*Professor, Department of Computer Science and Engineering,

Sri Ramakrishna Engineering College, Vattamalaipalayam, N.G.G.O Colony Post, Coimbatore-641022.

perumalpcse@gmail.com

²Associate Professor, Department of Computer Science and Engineering,

Sri Ramakrishna Engineering College, Vattamalaipalayam, N.G.G.O Colony Post, Coimbatore-641022.

Abstract: The automatic document clustering and topic extraction from the corpus provides a very essential requirement in many real time applications. The document clustering and topic detection is utilized to locating data quickly. Hence, in this paper, Type 2 Intuitionistic Fuzzy Clustering and Seagull Optimization Algorithm (Type 2 IFCSOA) is developed for document clustering and topic detection. The Type 2 IFCSOA is utilized to cluster the documents. Additionally, ensemble approach is utilized to identify by the topics from the clustered documents. In the proposed methodology, the pre-processing is utilized to remove unwanted information from the documents such as tokenization, stop word removal and stemming process. After that, the proposed method is utilized to cluster the documents. The clustered documents are labeled with the basis of clusters. After that, to achieve topic detection, the ensemble approach is utilized with feature extraction phases such as Term Frequency-Inverse Document Frequency (TF-IDF), Mutual information (MI), Text Rank Algorithm and analysis of keyword taking out from co-occurrence statistical -Information (CSI). The proposed methodology is implemented in MATLAB and performances were evaluated with the statistical measurements such as precision, recall, accuracy, sensitivity, purity measure and entropy. The proposed method is compared with the conventional methods such as Fuzzy C Means clustering (FCM), FCM-Particle Swarm Optimization (PSO), FCM-Genetic Algorithm (GA) and K means clustering.

Keywords: *document clustering, topic detection, feature extraction, term frequency, inverse document frequency and mutual information.*

1. Introduction

A large number of online scientific publications are written on the World Wide Web or on the Internet every day in the digital era [1]. There is generally increased interest in identifying topics in published scientific papers, including defining boundaries in the scientific field and identifying innovative capabilities, [2] which in turn sets out to be the central part of the process of producing scientific knowledge. Resource-rich research demands knowledge of earlier research and technology [3]. These will in turn affect subjects and form the foundation for advanced research based on earlier growths to achieve new findings and advances. Academic literature is generally a reliable data source [4]. The growth rate of knowledge and science has been examined through publications. Academic publications can make science communications possible. Written documents therefore include scientific and science explanations and make up the existing scientific literature [5].

Extraction of topic (TE), known as 'theme discovery,' 'theme identification,' 'theme discovery,' 'tag identification,' extraction of a keyword, category tagging, and cluster tagging' are all useful in a number of practical applications [6]. For example, by examining new publications on computer science, it is possible to identify increasingly important areas; in the foreseeable future, their trends and popularity may be further predicted [7]. There is however much uncertainty in terms of how these topics are defined, and there is a continuing discussion about how to automatically extract them. Furthermore, it is slower, costly, and error-prone to use manual methods to extract these subject areas [8, 9]. One of the most common techniques for identifying topics is to compile papers in order to recognize exact collections of academic papers that signify an associated topic. The most relevant terms are then extracted and classified from each cluster [10].

The Clustering of Text Documents (TDC) [11] is one of the strongly and effectively unmonitored text-mining learning techniques. From this method, the documents are clustered or grouped based on the similar keywords and placed in a comparable cluster. However, various documents are located in various clusters. Metaheuristic algorithms have made considerable efforts in recent years to solve the TDC problem [12]. This is because of the weaknesses of today's deterministic means of finding globally optimal solutions to address the problem. The majority of algorithms are based on the combined method (EA) [13], based on evolutionary algorithms, fittest theory survival and TA. The method of clustering documents is used [14]. Clustering methods have been improved with regard to the consideration of artificial (AI) [15] techniques such as Particle Swarm Optimization (PSO), Whale Optimization (WOA), and GWO. This paper has the main contributions,

The main contribution of the paper

The main contribution of the paper is presented as follows,

- In this paper, Type 2 IFCSOA is developed for document clustering and topic detection. The Type 2 IFCSOA is utilized to cluster the documents.
- Additionally, ensemble approach is utilized to identify by the topics from the clustered documents.
- In the proposed methodology, the pre-processing is utilized to remove unwanted information from the documents such as tokenization, stop word removal and stemming process. After that, the proposed method is utilized to cluster the documents. The clustered documents are labeled with the basis of clusters.
- To achieve topic detection, the ensemble approach is utilized with feature extraction phases such as TF-IDF, MI, Text Rank Algorithm and CSI.
- The proposed methodology is implemented in MATLAB and performances were evaluated with the statistical measurements such as precision, recall, accuracy, sensitivity, purity measure and entropy. The proposed method is compared with the conventional methods such as FCM, FCM-PSO, FCM-GA and K means clustering.

The remaining part of the paper is organized as follows, section 2 provides the review section of the topic detection and document clustering. The detail description of the proposed topic detection and document clustering is presented in the section 3. The result and discussion of the proposed methodology is presented in section 4. Conclusion of the paper is presented in section 5.

2. Literature Review

Many different methods are available to topic detection and document clustering. Some of the methods are reviewed in this section.

Stephan A. Curiskis *et al.*, [16] analyzed Embedding models for neural network document clustering and thematic modeling on 3 Twitter and Reddit datasets. Here, the benchmarking is to provide four different characteristics of term-frequency inversion-document-frequency (tf-idf) and word embedding models combined with four clustering methods. Several different measurement measures have been used in literature, thus giving the most appropriate extrinsic measures for this task a discussion and recommendation.

Peng Yang *et al.* [17] have developed a probabilistic tpLDA model, which incorporates different levels of information on the topic popularity to determine the prior distribution of the LDA, identify latent topics and improve clustering. In particular, the global topic popularity is introduced to reduce the potential distraction in the popularity of local clusters and local popularity focuses more on parts of the global theme. Both popularities

provide complementary information and can dynamically adjust the model's statistical parameters to their integration.

An ensemble method for automatic extraction of topics (ATT) was introduced in the paper by Kamal Abasi *et al.* [18], based on a series of scientific publications in text documents intended to extract topics from clustered documents. The current TE methods often draw on the theory of statistics. However, if the same clustered document is used, the results could be different. Consequently, inaccurate results, due to the behavior of TE methods, can be found in the topics extracted from the clustered papers.

In order to build effective model documents representation Nabil Alami *et al.* [19] have submitted an approach using clustering, topic modelling, and unattended neural networks. First, on a large text collection, a document grouping technique using an Extreme learning machine was performed. Second, topic modeling was used to collect documents, so that the subjects present in each cluster could be identified. Third, in a thematic space every document was represented by a matrix where the rows show the phrases of the document.

Di wu *et al.* [20] presented short text clustering algorithms for Micro-Blog Hot Topic Discovery based on BTM and GloVe similitudes of linear fusion (BG and SLF-Kmeans) The preprocessed microblog short texts have been used by BTM and GloVe. In order to calculate the similarity of the text using BTM topic modeling, the JS divergence was adopted. In calculating the text similarity based on GloVe word vector modeling, WMD was used for the improved word weight (IWMD). Finally, these two similarities are linearly fused and used to perform K-means clustering as the remote function.

3. Proposed System Model

Nowadays, the topic extraction and document clustering is an essential topic for many real-world applications. Moreover, there is uncertainty related to defining these topics, the efficient topic detection is designed. In this paper, Type 2 IFCSOA is developed to document clustering and topic detection. The complete architecture of the proposed system for document clustering and topic detection is presented in figure 1.

Figure. 1 Proposed System Architecture

Initially, the document database is collected from the open source system. The documents are separated with the specific topics. The collected documents are clustered and topics are identified with the assistance of the proposed methodology. After that, pre-processing is proceeding by tokenization, removal of stop words and stemming process. The pre-processed documents are utilized to clustered the documents. The clustered documents are labeled based on their clustered groups. The labeled documents are sent to the feature extraction phase with the ensemble approach for topic classification. The feature extraction of the document is achieved with the help of TF-IDF, MI, Text Rank Algorithm and CSI. Finally, the ensemble approach is utilized to classify the documents. The detail description of the proposed methodology is presented in below section.

3.1. Pre-processing

The text documents are presented in electronic format which increased in usage. Hence, the text clustering has been required technique to arrange the documents into clusters. The text clustering techniques mostly objective to create text papers clusters related to the papers with the basis of intrinsic contents. once start clustering method, the text documents should be procced with the pre-processing methods such as tokenization [21], removal of stop words and stemming [22] process. Hence, the text documents are changed into a required format.

3.2. Type-2 Intuitionistic Fuzzy C-Means Clustering for document clustering

The proposed clustering is utilized to topic detection and document clustering. The documents are clustered with the basis of the topics. The proposed clustering algorithm is utilized to cluster the documents based on the topics. The proposed clustering algorithm is an extension of Fuzzy C Means Clustering (FCM) and handles more uncertainty in data than FCM [23]. In the proposed methodology, two different fuzzifiers are utilized which are defined as a fuzzy degree. This proposed fuzzy degree is mathematically formulated as follows,

$$\begin{cases} J^1(U, V) = \sum_{i=1}^N \sum_{j=1}^C U_{ij}^{M^1}(X^i) D^2(X^i, V^j) \\ J^1(U, V) = \sum_{i=1}^N \sum_{j=1}^C U_{ij}^{M^2}(X^i) D^2(X^i, V^j) \end{cases} \quad (1)$$

Where, $D^2(X^i, V^j)$ can be described as Euclidean distance among the j^{th} cluster center and i^{th} pattern, $U_{ij}^{M^2}(X^i)$ can be described as the membership function of the i^{th} pattern related to the j^{th} cluster, C can be described number of clusters and N can be represented a number of data $X = \{X^1, X^2, \dots, X^N\}$. In the proposed clustering, the lower and upper membership functions are mathematically formulated as follows,

$$\overline{U}^{IJ}(X^i) = \max \left\{ \left[\sum_{K=1}^C \left(\frac{D(X^i, V^J)}{D(X^i, V^K)} \right)^{\frac{2}{M^1-1}} \right]^{-1}, \left[\sum_{K=1}^C \left(\frac{D(X^i, V^J)}{D(X^i, V^K)} \right)^{\frac{2}{M^2-1}} \right]^{-1} \right\} \quad (2)$$

$$\underline{U}^{IJ}(X^i) = \max \left\{ \left[\sum_{K=1}^C \left(\frac{D(X^i, V^J)}{D(X^i, V^K)} \right)^{\frac{2}{M^1-1}} \right]^{-1}, \left[\sum_{K=1}^C \left(\frac{D(X^i, V^J)}{D(X^i, V^K)} \right)^{\frac{2}{M^2-1}} \right]^{-1} \right\} \quad (3)$$

To compute the minimum value and maximum value of the j^{th} cluster center, the karnik mendal iterative algorithm is utilized. While the iterative algorithm is performed, the right memberships and left memberships are calculated related to each feature for a pattern which formulated as follows,

$$(U^{IJ})^L(X^i) = \frac{\sum_{L=1}^M U^{iL}}{M}, U^{iL} = \begin{cases} \overline{U}^{IJ}(X^i), & \text{if } X^{iL} \text{ uses } \overline{U}^{IJ}(X^i) \text{ for } V_j^L \\ \underline{U}^{IJ}(X^i), & \text{otherwise} \end{cases} \quad (4)$$

$$(U^{IJ})^R(X^i) = \frac{\sum_{L=1}^M U^{iL}}{M}, U^{iL} = \begin{cases} \overline{U}^{IJ}(X^i), & \text{if } X^{iL} \text{ uses } \overline{U}^{IJ}(X^i) \text{ for } V_j^R \\ \underline{U}^{IJ}(X^i), & \text{otherwise} \end{cases} \quad (5)$$

Where, M can be described as number of features of a pattern, V_j^L can be described as a minimum value of j^{th} cluster center, V_j^R can be described as a maximum value of j^{th} cluster center, $(U^{IJ})^R$ can be described as right membership functions and $(U^{IJ})^L(X^i)$ can be described as left membership function. The membership matrix and crisp centroids can be achieved with the consideration of defuzzification methods and type reduction methods.

$$V^J = \frac{V_j^L + V_j^R}{2} \quad (6)$$

$$U^{IJ}(X^i) = \frac{(U^{IJ})^L(X^i) + (U^{IJ})^R(X^i)}{2} \quad (7)$$

With the help of proposed clustering methods, the documents are clustered from the documents. The fuzzy membership function is an essential parameter to achieve the best cluster of the document process. The optimal fuzzy membership function is selected based on the SOA process. The detailed description of the proposed SOA-based fuzzy membership function selection is presented below section.

3.3. Seagull Optimization Algorithm

In this proposed methodology, the seagull optimization algorithm is developed to select optimal fuzzy membership function parameter in Type-2 Intuitionistic Fuzzy C-Means Clustering. The detail description of the seagull optimization algorithm is presented in this section. Seagulls are usually referred to as laridae and are a sea bird that can be shown throughout the world. There are many sea birds available that make sea-goats omnipresent and eat tuberculosis, amphibians, rodents, fish and eats. The seagulls are an intelligent bird [24], which draw fish with their feet by using crumbs that make rain. Fresh and salty water is eaten by the seagulls. The moose have a superior drum pair over their eyes, which is used to remove the salt from their system. The village is a most specified place for seagulls. Based on their own knowledge, the prey location is computed by seagulls. In the seagulls, the attacking mechanism and migration is a most significant process. Migration is defined as the seasonal changes of seagulls from one place to another place to compute most plentiful and richest energy sources. The mathematical model of attacking the prey and migration is presented as follows,

3.3.1. Migration

In the migration procedure, the algorithm proceeds with the seagulls transfer towards one position to another. In the migration process, the seagull must compensate three scenarios.

Avoiding the collisions: In the SOA, the collision should be avoiding among neighbors and additional parameters is described for the computation of optimal search agent position.

$$Cs = A \times Ps(X) \quad (8)$$

Where, A can be described as movement characteristics of search agent, X can be described as current iteration, Ps can be described as current position of search agent and Cs can be represented as search agent position that does not collision with remaining search agents. The movement characteristics of the search agent is presented as follows,

$$A = Fc - (X \times \left(\frac{Fc}{Maxiteration} \right)) \quad (9)$$

$$\text{Where, } X = 0, 1, 2, \dots, \text{max iteration} \quad (10)$$

Where, Fc can be described as value 2, A can be described as linearly reduced from Fc to 0, Fc can be described to control the frequency of considering parameter.

Movement towards best neighbors' direction:

Once avoiding the Collision among neighbors, the search agents are locating towards the movements of best neighbor.

$$Ms = B \times (Pb(X) - PS(X)) \quad (11)$$

Where, B can be described as randomizes that is responsible for efficient balancing among exploitation and exploration, $Pb(X)$ can be described best fit search agent, $PS(X)$ can be described as search agent and Ms can be described as search agent position.

$$B = 2 \times S = A^2 \times RD \quad (12)$$

Where, RD can be described as in the value generated from the range of $[0, 1]$.

Best search agent:

At last, the search agent position is updated based on the below equation,

$$DS = |Cs + Ms| \quad (13)$$

Where, DS can be described as distance among the best fit search agent and search agent.

3.3.2. Attacking of prey

The exploitation motive is exploiting the experience and history of the search procedure. In the migration process, the seagulls can change the attack angle condition [25]. Based on weight and wings, the seagulls are maintaining the altitude. During the prey attacking, the spiral movement characteristics can be occurred in the air. This spiral movement behavior can be presented as follows,

$$X' = R * \cos(K) \quad (14)$$

$$Y' = R * \sin(K) \quad (15)$$

$$Z' = R * K \quad (16)$$

$$R = U * e^{KV} \quad (17)$$

Where, e can be described as base of the natural logarithm, u and v can be described as spiral shape constants, k can be described as random number within the range of $0 \leq k \leq 2\pi$, and radius of individually try of the curved is denoted by R . The search agent updating process is computed by below equation,

$$Ps(X) = (DS * X' * Y' * Z') + Pbs(X) \quad (18)$$

Where, $Pbs(X)$ is described as the optimal solution and the position of other search agents can be updated. The flowchart of the SOA is presented in figure 2.

Figure. 2 Flowchart of the proposed methodology

Fitness function

Once the initial population is completed, the fitness function is computed. Based on the fitness function, the fuzzy membership function in fuzzy clustering. The fitness function is evaluated with the consideration of the distance value. Hence, the fitness function is formulated with the minimization of distance. The fitness function is achieved by selecting the optimal fuzzy membership function of clustering. The fitness function is mathematically formulated as follows,

$$FF = \text{Min}\{\text{Distance}\} \quad (19)$$

$$D(D1, D2) = (\sum_{I=1}^N |T1, I - T2, I|^2)^{1/2} \quad (20)$$

Where, T can be described as the text contains from the documents. $D1 = (T11, T12, \dots, T1N)$ and $D2 = (T21, T22, \dots, T2N)$ can be considered as the document 1 and 2 respectively. Based on the fitness function, the optimal fuzzy membership function is selected which are utilized to enhance the optimal clustering process. The proposed SOA is started with a random generated population. After that, the search agents are updated their

positions based on the optimal search agent under the iteration procedure. The A value is linearly reduced from the FC to 0. To achieve the smooth transition among exploitation and exploration, the variable B is responsible. So, the SOA can be considered as the global optimal solution because of their optimal exploitation and exploration capability. Once documents are clustered, the topic detection is proceeding with the feature extraction and ensemble method. The ensemble method is utilized to topic detection from the clustered documents.

3.4. Proposed Ensemble method for topic detection

Based on the proposed clustering algorithm, the labelling the cluster becomes essential so, the essential topics are detected to compute the clusters. Initially, the keyword extraction is essential requirement to topic classification. The RAKE [26] algorithm for automatic keyword extraction is used to compute the keyword of each cluster-related document for key frame extraction. For key frame extraction, Pre-processing steps like stop word removal, frequency breaks are used in the RAKE algorithm. The frequency cut is used to allow users to omit many less common terms. After that, it is sent to the feature extraction method for topic calculation. The feature extraction methods are explained in the following section.

3.4.1. Feature extraction

The feature extraction is an essential requirement to classify the documents based on their topics. In this proposed methodology, the MI, TextRank algorithm, CSI, and TF-IDF. The features are utilized to classify the documents. The feature extraction method mathematical computation is presented in this section.

MI

The MI is utilized to compute how much data in the absence or presence in a specific term to compute to determining the optimal clustering on a required cluster. Based on the concept of knowledge, the reciprocal information provided to reduce cluster uncertainty [27]. The reciprocal information between the exact word and group is presented as follows:

$$MI(T, C) = \sum_{I=0}^1 \sum_{J=0}^1 P(T^I, U^J) \log 2 \frac{P(T^I, U^J)}{P(T^1)P(U^J)} \quad (21)$$

Where, I can be described as specific term, C can be described as cluster, $P(T^1)$ and $P(U^J)$ can be described as probabilities which event occur independently, T^0 can be described as the term specified in Article does not include, or $T^0 U^0$ can be described as indicates a file which not related to the cluster. This probability value can be computed with the dividing the calculated events frequency by the documents.

TextRank Algorithm

Text rank algorithm can be represented as model that is graph based ranking for text mining. The TR algorithm can be utilized in many natural language processing methods which consisting of sentence and keyword

abstraction from the input of the documents. In the TR algorithm, the extract keywords are tokenized. The edge can then be generated between two nodes. In a precise word, the words were co-occurring using the window. This algorithm can be calculated in connection with the votes in each term. Each term score is computed as follows,

$$s(I) = (1 - D) + D \times \sum_{J \in IN(I)} \frac{1}{|Out(I)|} s(J) \quad (22)$$

Where, $Out(I)$ can be described as the set of vertices which term points, $IN(I)$ can be described as vertices set, D can be described as constant factor which is often to 0.85.

CSI

The CSI technique to compute significant terms in related with the word co-occurrence circulation, documents and alike documents distribution in cluster. Initially, the frequent terms are computed, co-occurrences can compute co-occurrences in an equal document among recurrent terms and collections of the terms [28]. The degree of probability distribution of frequent term is computed as follows,

$$\kappa^2(I) = \sum_{J=1}^M \frac{(FRE(i,J) - N_I P_J)^2}{N_I P_J} \quad (23)$$

Where, M can be described as frequent terms, N_I can be described as co-occurrence of total number, P_J can be described as frequent term of specific expected probability and κ^2 can be described as statistics-based keyword extraction method. Hence, the robustness is enhanced with the consideration of document clustering. To enhance the performance of statistics, the statistics can be formulated as follows,

$$\kappa'^2(I) = \kappa^2(I) - \max_{J \in M} \left(\frac{(FRE(i,J) - N_I P_J)^2}{N_I P_J} \right) \quad (24)$$

Normally, the CSI is proceeding with six phases. The first stage can be proceeding with stemming process in the text document. After that, Apriori algorithm is utilized to phrase extraction. Additionally, the frequent parameter is selected from the phrase extraction. Using the Jensen-Shannon divergence, the frequent term can be clustered. The probability of the document can be calculated. Finally, the statistics value can be keyword extraction based on above formulation.

TF-IDF

The TF-IDF shall be used to extract the documents that are considered a vector of TF-IDF weights from the text document. TF-IDF [29] the value of document term, the frequency product and the frequency of the opposite document term, the inverse frequency of the document. This condition can be formulated as follows,

$$TF - IDF(I) = TF(I) \times \left(\log \frac{M}{DF(J)} \right) \quad (25)$$

Where, $DF(j)$ can be described as quantity of documents which at least once consist the term j , M can be described as significance of complete documents-based number of clusters and $TF(i)$ can be described as frequency of term in the complete documents based on number of clusters. The complete features are combined; the topics are classified with the ensemble method. The outcomes of the proposed methodology are evaluated in below section.

5. Results and Discussion

The performance of the proposed methodology is validated in this section by simulation analysis. The main motive of the research is classifying the topic and document clustering from collected documents. The performance of the proposed methodology is evaluated and justified in this section. The proposed methodology is validated with the consideration of statistical measurements such as accuracy, precision, recall, sensitivity, specificity, F_Measure, purity measure and entropy respectively. The proposed methodology is compared with existing methods such as FCM-PSO, FCM-GA and K means clustering respectively. The implementation parameters of the proposed method are presented in table 1. The proposed method is validated by statistical measurements. The performance metrics of the statistical measurements are presented below in this section. To validate the performance of the proposed methodology, the related documents are collected from [30].

Dataset description:

20Newsgroups: This database is consisting of 19,997 articles which related with 20 classes. This database is collected from various Usenet newsgroups. In this proposed methodology, the initial 100 documents are selected from the most frequent topic classes such as rec autos classes, talk politics misc. and comp windows.

Table. 1 Implementation parameters of the proposed methodology

S. No	Method	Description	Value
1	Proposed methodology	Number of clusters	5
2		Stop parameter	0.1
3		Size of initial population	10
4		Mutation frequency	0.5
10	Seagull optimization algorithm	Population size	50
11		Lower bound	10
12		Upper bound	-10
13		Maximum iteration	200
14		population male	20
15		Population female	20
16		Inertia weight	0.8
17		Inertia weight damping ratio	1
18		Personal learning coefficient	1.0
19		Global learning coefficient	1.5

The proposed method is used to topic detection and document clustering. The confusion matrix is computed based on following constraints,

- A topic is actually presented and detected as presented scenario which named as True Positive (TP).
- A topic is actually not presented and detected as not presented which named as True Negative (TN).
- A topic is actually not presented but detected which named as False Positive (FP).
- A topic is presented and detected as not presented which named as False Negative (FN).

Based on the progress of confusion matrix terms, the proposed methodology is evaluated by performance metrics which are formulated as follows,

Accuracy: It is defined as number of correctly detected data instances from the total number of instances.

The formula of accuracy is presented follows,

$$Accuracy = \frac{TN+TP}{TN+FP+TP+FN} \quad (26)$$

Specificity: It is defined as fraction of positive detected which actually correct. The precision effectively computed a complete probability and performance measure. The specificity formulation presented follows,

$$specificity = \frac{TP}{TP+FP} \quad (27)$$

Recall: It can be defined as a ratio of positive samples correctly detected to positive total instances, as set out below.

$$Recall = \frac{TP}{(TP+FN)} \quad (28)$$

Sensitivity: It can be defined as the proportion of negative instances that have correctly been detected in the following total negative instances;

$$sensitivity = \frac{TN}{(TN+FP)} \quad (29)$$

F_Measure: The F_Measure must be presented among 0 to 1. The worst value is 0 and best value is 1.

The F_Measure can be represented as follows,

$$F_{Measure} = \frac{2TP}{(2TP+FP+FN)} \quad (30)$$

Purity Measure:

The purity measure is used to calculate the percentage of documents and objects assigned to each cluster in the most common class. The effective value of pureness is equal to 1. The measurement of purity can be calculated using the following equation,

$$Purity = \frac{1}{N} \sum_{I=1}^K MAX(I, J) \quad (31)$$

Where, K can be described as quantity of clusters, N can be described as whole number of documents in the dataset, $MAX(I, J)$ can be represented as large class size I with the cluster J .

Entropy:

The entropy test measures are computed to identify the correct clusters for every class. The entropy measure is computed by below equation,

$$E(KJ) = -\sum_{I=1} P(I, J) \log(P^I, J) \quad (32)$$

Where, probability of the clusters within the documents are represented as $P(I, J)$, entropy of the cluster is represented as $E(KJ)$. The complete clusters, the entropy is computed as follows,

$$Entropy = -\sum_{I=1} \frac{NI}{N} E(KJ) \quad (33)$$

Where, K can be described as clusters, NI can be described as quantity of papers in the cluster, N can be described as the quantity of documents presented in the datasets.

Case 1: Document Clustering

In this section, the proposed document clustering is evaluated and compared with the conventional methods by statistical measurements. To evaluate the proposed methodology, 100 documents are collected and proceed with the proposed methodology.

Figure. 3 Analysis of accuracy

Figure. 4 Analysis of Recall

Figure. 5 Analysis of specificity

Figure. 6 Analysis of sensitivity

The proposed method is validated by using comparison analysis. In the comparison analysis, the proposed method is compared by proposed, FCM-PSO, FCM-GA and k means clustering respectively. The comparison analysis of accuracy is presented in figure 3. From the figure 3, the proposed method is achieved the 0.78 rate accuracy in document clustering. Similarly, the FCM-PSO, FCM-GA and k means clustering methods are achieved the accuracy 0.75, 0.68 and 0.62 respectively. From the comparison analysis, the proposed method is achieving the best accuracy in document clustering. The comparison analysis of recall is presented in figure 4. From the figure 4, the proposed method is achieved the 0.79 rate recall in document clustering. Similarly, the FCM-PSO, FCM-GA and k means clustering methods are achieved the recall 0.71, 0.68 and 0.62 respectively. From the comparison analysis, the proposed method is achieving the best recall in each cluster in document clustering. The comparison analysis of specificity is presented in figure 5. From the figure 5, the proposed method is achieved the 0.72 rate precision in document clustering. Similarly, the FCM-PSO, FCM-GA and k means clustering methods are achieved the specificity 0.62, 0.60 and 0.58 respectively. From the comparison analysis, the proposed method is achieving the best specificity in every cluster in document clustering.

Figure. 7 Analysis of F_Measure

Figure. 8 Analysis of Purity measure

Figure. 9 Analysis of entropy

The comparison analysis of sensitivity is presented in figure 6. From the figure 6, the proposed method is achieved the 0.78 rate sensitivity in document clustering. Similarly, the FCM-PSO, FCM-GA and k means clustering methods are achieved the sensitivity 0.71, 0.62 and 0.60 respectively. From the comparison analysis, the proposed method is achieving the best sensitivity in document clustering. The comparison analysis of F_Measure is presented in figure 7. From the figure 7, the proposed method is achieved the 0.73 rate F_Measure. Similarly, the FCM-PSO, FCM-GA and k means clustering methods are achieved the F_Measure 0.68, 0.61 and 0.59 respectively. From the comparison analysis, the proposed method is achieving the best F_Measure in document clustering. The comparison analysis of purity measure is presented in figure 8. From the figure 8, the proposed method is achieved the 0.72 rate purity measure. Similarly, the FCM-PSO, FCM-GA and k means clustering methods are achieved the purity measure 0.62, 0.60 and 0.59 respectively. From the comparison analysis, the proposed method is achieving the best purity measure in document clustering. The comparison analysis of entropy is presented in figure 9. From the figure 9, the proposed method is achieved the 0.79 rate entropy measure. Similarly, the FCM-PSO, FCM-GA and k means clustering methods are achieved the entropy measure 0.75, 0.68 and 0.62 respectively. From the comparison analysis, the proposed method is achieving the best entropy measure in document clustering.

Case 2: Topic Detection

In this section, the proposed ensemble approach-based topic detection is evaluated and compared with the conventional methods. The three different topics are classified with the help of proposed methodology. The proposed methodology has been validated in this section by statistical measurements.

Figure. 10 Analysis of accuracy

Figure. 11 Analysis of Recall

Figure. 12 Analysis of specificity

Figure. 13 Analysis of sensitivity

The proposed method is validated by using comparison analysis. In the comparison analysis, the proposed method is compared by proposed, FCM-PSO, FCM-GA and k means clustering respectively. The comparison analysis of accuracy is presented in figure 10. From the figure 10, the proposed method is achieved the 0.81 rate accuracy in document clustering. Similarly, the FCM-PSO, FCM-GA and k means clustering methods are achieved the accuracy 0.73, 0.65 and 0.60 respectively. From the comparison analysis, the proposed method is achieving the best accuracy in document clustering. The comparison analysis of recall is presented in

figure 11. From the figure 11, the proposed method is achieved the 0.80 rate recall in document clustering. Similarly, the FCM-PSO, FCM-GA and k means clustering methods are achieved the recall 0.75, 0.71 and 0.65 respectively. From the comparison analysis, the proposed method is achieving the best recall in each cluster in document clustering. The comparison analysis of specificity is presented in figure 12. From the figure 12, the proposed method is achieved the 0.74 rate precision in document clustering. Similarly, the FCM-PSO, FCM-GA and k means clustering methods are achieved the specificity 0.65, 0.63 and 0.61 respectively. From the comparison analysis, the proposed method is achieving the best specificity in every cluster in document clustering. The comparison analysis of sensitivity is presented in figure 13. From the figure 13, the proposed method is achieved the 0.79 rate sensitivity in document clustering. Similarly, the FCM-PSO, FCM-GA and k means clustering methods are achieved the sensitivity 0.74, 0.68 and 0.65 respectively. From the comparison analysis, the proposed method is achieving the best sensitivity in document clustering.

Figure. 14 Analysis of F_Measure

Figure. 15 Analysis of Purity measure

Figure. 16 Analysis of entropy

The comparison analysis of F_Measure is presented in figure 14. From the figure 14, the proposed method is achieved the 0.76 rate F_Measure. Similarly, the FCM-PSO, FCM-GA and k means clustering methods are achieved the F_Measure 0.71, 0.65 and 0.63 respectively. From the comparison analysis, the proposed method

is achieving the best F_Measure in document clustering. The comparison analysis of purity measure is presented in figure 15. From the figure 15, the proposed method is achieved the 0.74 rate purity measure. Similarly, the FCM-PSO, FCM-GA and k means clustering methods are achieved the purity measure 0.65, 0.62 and 0.61 respectively. From the comparison analysis, the proposed method is achieving the best purity measure in document clustering. The comparison analysis of entropy is presented in figure 16. From the figure 16, the proposed method is achieved the 0.81 rate entropy measure. Similarly, the FCM-PSO, FCM-GA and k means clustering methods are achieved the entropy measure 0.76, 0.64 and 0.61 respectively. From the comparison analysis, the proposed method is achieving the best entropy measure in document clustering.

6. Conclusion

In this paper, Type 2 IFCSOA has been developed for document clustering and topic detection. The Type 2 IFCSOA has been utilized to cluster the documents. Additionally, ensemble approach has been utilized to identify by the topics from the clustered documents. In the proposed methodology, the pre-processing has been utilized to remove unwanted information from the documents such as tokenization, stop word removal and stemming process. After that, the proposed method has been utilized to cluster the documents. The clustered documents have been labeled with the basis of clusters. After that, to achieve topic detection, the ensemble approach has been utilized with feature extraction phases such as TF-IDF, MI, Text Rank Algorithm and CSI. The proposed methodology is implemented in MATLAB and performances were evaluated with the statistical measurements such as precision, recall, accuracy, sensitivity, purity measure and entropy. The proposed method is compared with the conventional methods such as FCM, FCM- PSO, FCM- GA and K means clustering. From the outcome analysis, the proposed methodology has been achieved best results in terms of accuracy. In future, the large amount of database with various topics will be analyzed with effective methods.

Data Availability Statement

Not Applicable.

References

- [1]. Reena, Pagare, Akhil Khare, and Shankar Chaudhary, "(T-ToCODE): A Framework for Trendy Topic Detection and Community Detection for Information Diffusion in Social Network", In Data Management, Analytics and Innovation, (2019) 1024:613-632.

- [2]. Jagrati Singh, and Anil Kumar Singh, "NSLPCD: Topic based tweets clustering using Node significance based label propagation community detection algorithm", *Annals of Mathematics and Artificial Intelligence*, (2021) 89(3-4):371-407.
- [3]. Shaymaa H.Mohammed, and Salam Al-augby, "LSA & LDA Topic Modeling Classification: Comparison study on E-books", *Indonesian Journal of Electrical Engineering and Computer Science*, (2020) 19(1):353-362.
- [4]. Di Wu, Ruixin Yang, and Chao Shen, "Sentiment word co-occurrence and knowledge pair feature extraction based LDA short text clustering algorithm", *Journal of Intelligent Information Systems*, (2021) 56(1):1-23.
- [5]. Junya Tang, Li Li, Ying Liu, and Kuo-yi Lin, "Automatic Identification of Bottleneck Tasks for Business Process Management Using Fusion-based Text Clustering", (2021):1-6
- [6]. Mohamed Aymen Ben HajKacem, Chiheb-Eddine Ben N'cir and Nadia Essoussi, "A parallel text clustering method using Spark and hashing", *Computing* (2021): 1-25.
- [7]. Karpagalingam Thirumoorthy, and Karuppaiah Muneeswaran, "A hybrid approach for text document clustering using Jaya optimization algorithm", *Expert Systems with Applications*, (2021) 178:115040.
- [8]. Caiquan Xiong, Zhen Hua, Ke Lv, and Xuan Li, "An Improved K-means text clustering algorithm by Optimizing initial cluster centers", *Proceedings - 2016 7th International Conference on Cloud Computing and Big Data, CCBD 2016* (2017) 265-268.
- [9]. Laith Mohammad Abualigah, and Ahamad Tajudin Khader, "Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering", *Journal of Supercomputing* (2017) 73(11) 4773-4795.
- [10]. R.Janani and S. Vijayarani, "Text document clustering using spectral clustering algorithm with particle swarm optimization", *Expert Systems with Applications*, (2019) 134:192-200.
- [11]. Mingyu Yao, Dechang Pi and Xiangxiang Cong, "Chinese text clustering algorithm based k-means", *Physics Procedia*, (2012) 33:301-307.
- [12]. Jung-Yi Jiang, Ren-Jia Liou, and Shie-Jue Lee, "A fuzzy self-constructing feature clustering algorithm for text classification", *IEEE Transactions on Knowledge and Data Engineering*, (2011) 23(3):335-349.
- [13]. Andrew Skabar and Khaled Abdalgader, "Clustering sentence-level text using a novel fuzzy relational clustering algorithm", *IEEE Transactions on Knowledge and Data Engineering*, (2013) 25(1):62-75.

- [14]. Abasi, Ammar Kamal, Ahamad Tajudin Khader, Mohammed Azmi Al-Betar, Syibrah Naim, Sharif Naser Makhadmeh, and Zaid Abdi Alkareem Alyasseri, "Link-based multi-verse optimizer for text documents clustering", *Applied Soft Computing*, (2020) 87:106002.
- [15]. Laith Abualigah, Amir H. Gandomi, Mohamed Abd Elaziz, Husam Al Hamad, Mahmoud Omari, Mohammad Alshinwan, and Ahmad M. Khasawneh, "Advances in meta-heuristic optimization algorithms in big data text clustering", *Electronics (Switzerland)*, (2021) 10(2):1-29.
- [16]. Stephan A. Curiskis, Barry Drake, Thomas R. Osborn and Paul J. Kennedy, "An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit", *Information Processing and Management*, (2020) 57(2):102034.
- [17]. Peng Yang, Yu Yao and Huajian Zhou, "Leveraging global and local topic popularities for LDA-based document clustering", *IEEE Access*, (2020) 8:24734-24745.
- [18]. Ammar Kamal Abasi, Ahamad Tajudin Khader, Mohammed Azmi Al-Betar, Syibrah Naim, Zaid Abdi Alkareem Alyasseri and Sharif Naser Makhadmeh, "An ensemble topic extraction approach based on optimization clusters using hybrid multi-verse optimizer for scientific publications", *Journal of Ambient Intelligence and Humanized Computing*, (2021) 12(2):2765-2801.
- [19]. Nabil Alami, Mohammed Meknassi, Noureddine En-nahnahi, Yassine El Adlouni and Ouafae Ammor, "Unsupervised neural networks for automatic Arabic text summarization using document clustering and topic modeling", *Expert Systems with Applications*, (2021) 172:114652.
- [20]. Di Wu, Mengtian Zhang, Chao Shen, Zhuyun Huang and Mingxing Gu, "Btm and glove similarity linear fusion-based short text clustering algorithm for microblog hot topic discovery", *IEEE Access*, (2020) 8:32215-32225.
- [21]. Samer Abdulateef, Naseer Ahmed Khan, Bolin Chen and Xuequn Shang, "Multidocument Arabic text summarization based on clustering and Word2Vec to reduce redundancy", *Information (Switzerland)*, (2020) 11(2):59.
- [22]. Tinghuai Ma, Raeed Al-Sabri, Lejun Zhang, Bockarie Marah and Najla Al-Nabhan, "The impact of weighting schemes and stemming process on topic modeling of Arabic long and short texts", *ACM Transactions on Asian and Low-Resource Language Information Processing* (2020) 19(6):1-23.
- [23]. Imo Eyoh, Jeremiah Eyoh and Ini Umoeka, "Interval Type-2 Intuitionistic Fuzzy Logic System for Forecasting the Electricity Load", *International Journal of Advances in Scientific Research and Engineering*, (2020) 6(10):38-51.

- [24]. Natee Panagant, Nantiwat Pholdee, Sujin Bureerat, Khon Kaen, Ali Rıza Yıldız and Sadiq M. Sait, "Seagull optimization algorithm for solving real-world design optimization problems", *Materialpruefung/Materials Testing*, (2020) 62(6):640-644.
- [25]. Gaurav Dhiman, Krishna Kant Singh, Mukesh Soni, Atulya Nagar, Mohammad Dehghani, Adam Slowik, Amandeep Kaur, Ashutosh Sharma, Essam H. Houssein and Korhan Cengiz, "MOSOA: a new multi-objective seagull optimization algorithm", *Expert Systems with Applications*, (2021) 167: 114150.
- [26]. Engel. D, Rose. S, Cramer. N and Cowley. W, "Automatic keyword extraction from individual documents", *Text Mining: Applications and Theory*, (2010) 1:1-20.
- [27]. Koopman. R and Wang. S, "Mutual information based labelling and comparing clusters", *Scientometrics*, (2017) 111(2):1157-1167.
- [28]. Onan A, Korukoglu S and Bulut H, "Ensemble of keyword extraction methods and classifiers in text classification", *Expert Systems with Applications*, (2016) 57:232-247.
- [29]. Abasi, Ammar Kamal, Ahamad Tajudin Khader, Mohammed Azmi Al-Betar, Syibrah Naim, Sharif Naser Makhadmeh and Zaid Abdi Alkareem Alyasseri, "A novel ensemble statistical topic extraction method for scientific publications based on optimization clustering", *Multimedia Tools and Applications*, (2021) 80(1):37-82.
- [30]. <https://www.kaggle.com/crawford/20-newsgroups>