Journal of Information Systems Engineering and Management
2025, 10(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Spectral Clustering-Based Particle Swarm Optimization Algorithm for Document Clustering

T. Elavarasi1*, Dr. R. Nagarajan2
*1Research scholar, Department of computer and information science, Annamalai University, Annamalainagar, Tamil Nadu, India.
2Assistant Professor/Programmer, Department of computer and information science,
Annamalai University, Annamalainagar, Tamil Nadu, India.
*Corresponding Author E-Mail: arasithirugnanam@gmail.com

ARTICLE INFO

Received: 11 Oct 2024
Revised: 10 Dec 2024
Accepted: 22 Dec 2024

ABSTRACT

The process of automatically grouping documents into clusters such that the documents in one cluster are very comparable to the documents in the remaining clusters have been known as document clustering. Due to its broad application in a number of fields, including search engines, web mining, and information retrieval, it has been the subject of much research. It involves clustering documents that are identical to one another and calculating how identical they are. It facilitates simple navigation by offering effective document representation as well as visualization. Hence, this research paper plans to perform the document clustering using the nature inspired optimization technique. Initially, the dataset is manually gathered from different sources. Next, the data preparation has been done for extracting the text content from the published documents. These prepared data undergo pre-processing for removing the punctuations, stop words, and lowercase conversion. The features are extracted from these pre-processed data utilizing the Term Frequency-Inverse Document Frequency (TF-IDF) approach for extracting the keywords. The extracted features undergo the final clustering phase employing the spectral clustering algorithm, in which the parameter tuning has been done by the nature inspired optimization algorithm referred as Particle Swarm Optimization (PSO) with the consideration of silhouette score maximization as the objective function. This proposed spectral clustering-PSO clusters the final output into six classes such as data mining, deep learning, image, machine learning, network, and sports respectively. The proposed document clustering model describes its betterment over the remaining techniques with respect to distinct measures. The proposed spectral clustering-PSO in terms of silhouette score is 51.92%, 70.81%, 45.93%, and 20.89% better than JA-GWO, tpLDA, HDMA, and Net2Vec respectively. Similarly, the proposed spectral clustering-PSO in terms of davies bouldin score is 89.69%, 58.48%, 32.67%, and 13.99% advanced than JA-GWO, tpLDA, HDMA, and Net2Vec respectively.

Keywords: Document Clustering; Spectral Clustering; Particle Swarm Optimization; Term Frequency-Inverse Document Frequency.

1. INTRODUCTION

Massive amounts of digital data are saved and retrieved daily via smart devices on the internet, in which billions of people have access to it through the proliferation of smart apps [1]. Text analysis has been crucial in the area of text mining since there exists an excess of e-data; managing a large number of documents requires specialized techniques. Document clustering describes an effective technique for grouping a collection of ambiguous documents into a subset of related as well as comparable clusters in the fields of pattern recognition, text mining, and machine learning [2]. A subject is often described as a set of phrases with statistically important semantic associations in document clustering research [3]. An email, a book chapter, a blog post, a journal article, or any other kind of unstructured text may all be considered documents. Labels or any earlier understanding of the text are not necessary for this method to function. The method of clustering involves assembling similar-behaving items into homogeneous groups [4].

Feature selection have been often used to obtain the best and major beneficial characteristics for documents. However, there exists still a significant difficulty with huge dimensionality in documents [5]. The document clustering becomes complex and the effectiveness suffers from big dimensions when the document collection has numerous hundred text characteristics. Additionally, the accuracy related to the resulting clusters decreases

with an increase in text dimensionality [6]. Therefore, in order to eliminate superfluous and unnecessary text characteristics from documents, feature selection is required. The aforementioned issues have been resolved recently, and deep learning approaches have demonstrated impressive performance in examining the semantic description of texts and words. Neural Networks (NNs) are designed to frequently describe their weight values and take on tasks like document clustering using layer-wise pre-training [7]. On the other hand, an overabundance of hyperparameters describes the primary issue with deep learning.

Furthermore, the development of metaheuristic optimization methods in modern times aids in the resolution of numerous problems pertaining to various domains, including text document clustering, image processing, data mining, etc. [8]. These methods nevertheless have to contend with the problem of rapid convergence or premature. However, these optimization strategies may often be part of the local optima issue [9]. Even though several research groups have suggested optimization methods for the text feature selection procedure, the high dimension associated with these methods causes performance degradation as they focus on selecting a novel subset of text characteristics utilizing the earlier acquired feature subset [10]. An effective as well as promising solution has to be suggested to remove these issues in document clustering. As a result, the need for creative feature selection employing optimization algorithm methods has been taken into account and inspired.

The paper contribution is as below.

- To perform the document clustering using the nature inspired optimization technique by manually gathering the dataset from different sources.
- To do the data preparation for extracting the text content from the published documents and it undergoes pre-processing for removing the punctuations, stop words, and lowercase conversion.
- To extract the features using the TF-IDF for extracting the keywords.
- To undergo the final clustering step using the spectral clustering algorithm, in which the parameter tuning has been done by the nature inspired optimization algorithm referred as PSO with the consideration of silhouette score maximization as the objective function.
- To cluster the final output into six classes such as data mining, deep learning, image, machine learning, network, and sports by the proposed spectral clustering-PSO method.

The paper organization is as follows. Section 1 describes the introduction of document clustering. Section 2 elaborates literature survey. Section 3 explains proposed methodology with proposed model, dataset collection, data preparation, pre-processing, feature extraction using TF-IDF, clustering using spectral clustering, and PSO algorithm. Section 4 evaluates results. Section 5 ends with conclusion.

1.1 Motivation

One common datamining usage in search engines has been document clustering. The quantity of documents has been steadily rising. The volume of electronic documents has been growing, making it more difficult to manually arrange, evaluate, and present them effectively. While people are capable of identifying clusters in two and three dimensions, high-dimensional information requires algorithms when there exists a vast volume of information. Data has to be arranged somehow so that the needed documents are accessible and readily identified. Thus, there exists a requirement for text documents to be automatically grouped in an efficient and effective manner. Word occurrence in every document set determines whether documents should be grouped together. Document sorting, data visualization, document retrieval, document tag clustering, document analysis, and other applications may all benefit from document clustering. Examining many articles about document clustering not only provides useful information but also aids in identifying new problems in the field of clustering. There exist several approaches for solving the issue of document clustering.

2. RELATED WORK

With the usage of feature sets, namely TF-IDF, this study aimed to offer an optimal clustering method for text documents [11]. In this case, the first centroid pick was highly concentrated, allowing the suggested clustering procedure to frequently group the text utilizing a weighted similarity metric. To decrease weighted similarity within the documents, the weighted similarity function really takes into account the inter- as well as intra-cluster similarity of both sorted as well as unordered documents. The hybrid optimization method, which combined the

Grey Wolf Algorithm (GWO) and Jaya Algorithm (JA) suggested an enhanced clustering method; thus, the suggested approach was known as JA-oriented GWO.

A probabilistic model called tpLDA was put out [12]; it used various degrees of topic popularity data to ascertain the previous LDA distribution, find latent topics, as well as improve clustering. In particular, the introduction of global subject popularity served to mitigate any possible distractions from local cluster popularity, while it highlighted specific aspects related to the worldwide topic popularity. The two popularities provide complimentary data, and when they are combined, the statistical parameters associated with the method may be dynamically adjusted. Experiments conducted on real data sets demonstrate that the suggested methodology significantly increases the accuracy of document clustering when compared to the traditional methods.

The Elbow technique as well as the Silhouette score for cluster k identification was developed [13]. In order to describe text documents, this work primarily proposed three unique method combinations. On the document clustering issue, the suggested enhancements have been verified. The results were presented using a cluster analysis on the basis of two assessment measures: external as well as internal. The four alphanumeric datasets (Web KB, Reuters, Doc50, and News20) and the two numeric datasets (Wine and Iris) have each been the subject of experiments.

A brand method known as the Hierarchical Dirichlet Multinomial Allocation (HDMA) method was created for document clustering [14]. A hierarchical topic generating approach consisting of two steps was used to study the HDMA framework. While preserving the local peculiarities associated with every data source, topics were taught to share their generic properties across data sources. To determine the source-level subject focus, an exclusive topic partition was applied to every data source. The count of clusters for every data source and the HDMA method's parameters are next simultaneously discovered using a Gibbs sampling technique. The HDMA method's effectiveness was demonstrated by the outcomes of the experiments.

On the basis of a text classifier, an automatic consensus building metric was suggested [15]. Two different DIMs created the fundamental base clusters and document partitions. In order to identify concordant documents and provide a dataset for the text classifier to be trained, the consensus building metric approach made usage of the cluster data. The classifier created novel groups by predicting labels. To evaluate the effectiveness associated with the suggested method, eight standard data sets were used for the testing. The suggested consensus clustering's superiority over individual findings was demonstrated by the enhancement that was seen upon using it. The two discrimination information measures that we employed in the experiments to get the base clustering solutions were Measurement of Discrimination Information (MDI) and Relative Risk (RR).

For fine-grained document clustering, a probabilistic network graph was established, and a probabilistic generative method and computation technique were created [16]. Additionally, a unique NN-oriented network embedding learning technique was developed that represented the strength related to the association among the documents and takes into account the relevance of a document on the basis of its rankings. Reputable papers could be centralized displayed by taking into account the relevance of each document. The suggested ranking-oriented network-embedding technique outperformed the current network embedding techniques in assessment tests on a number of algorithms, including the typical word/phrase-oriented clustering techniques and k-means algorithm.

The texts were clustered using the Salp Swarm Algorithm (SSA) [17]. By using similarity as well as distance-oriented measures as the goal function in the clustering area, the research was enhanced. The purpose related to the experimental validation was to demonstrate the effectiveness associated with the SSA-oriented similarity distance measurement, which significantly enhanced the text document clustering process. The suggested SSA provided superior text document clustering with consideration to accuracy, specificity, sensitivity, and f-measure when compared to current techniques.

The best cluster centroids for K-means clustering were created using the Squirrel Search algorithm [18]. The SSC method, which was population-oriented and motivated by nature, was employed to explore the search space and identify the best initial cluster centres. The suggested method's efficacy was examined using nine distinct benchmark text datasets, including webpages, medline, and technical reports. The outcomes of the suggested technique were contrasted with those of the following algorithms: K-means, genetic, partial swarm optimization, harmonic search, and magnetic optimization.

For Document Clustering, a Deep Structural Enhanced network called DSEDC was developed [19]. For improved clustering efficiency, the DSEDC method augmented the AE with GCN. A complete document description, obtained by combining document external and internal semantics were learnt in an ensemble-reinforced enhancement scheme. The developed DSEDC technique outperformed the conventional deep document clustering methods, as shown by vast trials.

In order to achieve a higher clustering quality, a concept-oriented semi-supervised methodology was presented for document clustering [20]. This methodology made usage of both labelled as well as unlabeled data. A group of words with comparable semantic properties make up concepts. To use document labels for the purpose of extracting more pertinent ideas, the idea of semi-supervised concepts was introduced. Additionally, a fresh approach was provided to document clustering that utilized the weights associated with these ideas. The texts were described in the first as well as second stages of the suggested methodology using the ideas that were taken from the corpus's collection of embedded words. The suggested description maintained the document proximity data while being comprehensible. Unlabeled information was used to capture the entire format.

2.1 Problem definition

Organizing related documents into clusters have been known as document clustering. It describes a tried-and-true method for grouping a huge number of documents for analysis and insight without utilizing an expensive, prefabricated learning set. Especially in the current day, when millions of digital documents are generated daily, efficient as well as quick document clustering techniques are crucial. One among the biggest issues with clustering is that illegal nodes, like keyword stuffing or spam—that is, purposeful link manipulation or the addition of well-known terms may be seen as nodes in the clustering findings. Accurately determining the materials' importance defines the major crucial stage. In order to do this, a number of previous research have involved utilizing a platform like a search engine or analyzing the data found in as well as out of links to rank documents. Since these techniques rely on internal data that cannot be independently verified, they are susceptible to misuse. By identifying groups with a high degree of similarity on the basis of the largest number of comparable terms found in the documents, clustering methods are applied. Without any prior information, a person can obtain a strong understanding of a data set (main characteristics) by cluster analysis. As most clustering techniques need a high count of input parameters, cluster analysis seems to be often difficult.

3. PROPOSED METHODOLOGY

3.1 Proposed Model

The proposed document clustering model includes various phases such as data collection, data preparation, pre-processing, feature extraction, and clustering. The dataset is first collected manually from various sources. The next step involves preparing the data so that the text content of the published documents can be extracted. Pre-processing remains done on this prepared data to eliminate stop words, lowercase conversion, and punctuations. The TF-IDF approach stays utilized to extract the features from the pre-processed data by extracting the keywords. The extracted features are subjected to the last stage of clustering using the spectral clustering algorithm. The objective function of this step is silhouette score maximization, and the parameters are tuned using PSO, a nature-inspired optimization algorithm. The final result is clustered into six groups using the suggested spectral clustering-PSO, including data mining, deep learning, image, machine learning, network, and sports. The proposed published document clustering model has been shown in Fig. 1.

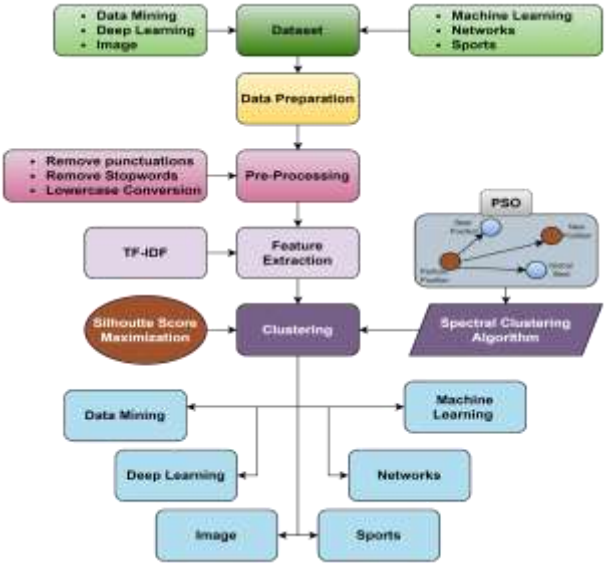
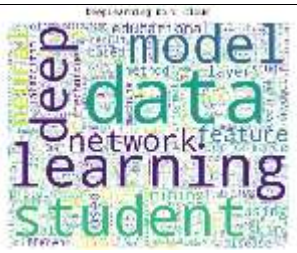


Fig. 1. Proposed Published Document Clustering Model

3.2 Dataset collection

The dataset for the proposed published document clustering model is gathered in terms of six manual sources such as data mining, deep learning, image, machine learning, network, and sports. Each source is composed of 10 published documents. These sources can be represented symbolically in the form of word cloud as shown in Fig. 2.

Dataset	Symbolic representation
Data mining	
Deep learning	
Image	

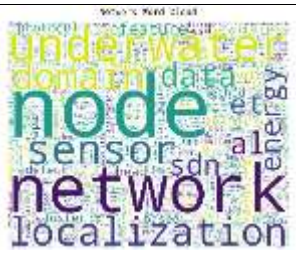
Machine learning	
Network	
Sports	

Fig. 2. Symbolic representation of word cloud for the gathered data sources

3.3 Data preparation

The data preparation is done in the proposed published document clustering model for extracting the text content from the published documents. It describes the technique of getting unprocessed data ready for additional handling and examination. With data, it assists in locating any trouble spots that could otherwise go unnoticed. It represents a crucial phase in the data analysis process. Although there exists a wealth of low-quality data on the Internet and in other data sources, it is crucial to understand how to clean the data so that it is usable for a variety of applications. This include fixing any mistakes, organizing your material, and finally formatting it appropriately.

3.4 Pre-processing

The prepared data of the proposed published document clustering model undergoes the pre-processing step for removing the punctuations, stop words, and lowercase conversion process.

Punctuation removal: In order to make the text simpler and more focused on the words itself, punctuation removal describes a text preprocessing step in which all punctuation marks—like commas, periods, emojis, exclamation marks, etc.—are removed. It aids in text simplification and cleanliness by getting rid of superfluous symbols. Generally speaking, it represents a good idea to remove punctuation before performing more research. This defines the technique of fixing mistakes or misspellings in written content. Correcting spelling in writing enhances its readability and quality and helps to prevent misunderstandings. This tool removes typographical signs like question marks, brackets, apostrophes, commas, colons, ellipses, dashes, periods, exclamation points, and other characters from strings that are loaded. The result represents a sanitized string that only contains alphanumeric characters. It streamlines the dataset, cutting down on complexity and enabling methods to concentrate on the text's semantic meaning.

Stop words removal: High frequency words with minimal semantic weight are known as stop words, and they are therefore unlikely to aid in retrieval. Eliminating stopwords makes the text smaller and less distracting, allowing readers to concentrate on the key words. The concept has been as simple as eliminating terms that appear often in every text in the corpus. Pronouns as well as articles are often categorized as stop words. When starting to utilize longer word sequences as method features, eliminating these stop words becomes much

more valuable. Because stop words are too common and don't contribute any useful data, eliminating them from the corpus reduces the amount of information and speeds up text analysis. The below are certain advantages of stop word removal: It may be easier to handle and analyze text information more quickly if its size have been decreased. lowering the quantity of meaningless words that Natural Language Processing (NLP) systems must parse in order to improve their effectiveness.

Lowercase conversion: Readability tends to be increased when user manuals as well as documentation are written in lowercase. Users will find it simpler to comprehend the material and follow directions without getting sidetracked by overuse of capitalization. Using lowercase letters may give writing a lot more of a relaxed, informal feel. Capital letters are perceived as hostile by certain readers. Text frequently uses a range of caps to emphasize sentence beginnings or proper nouns. For ease, it's usual practice to convert everything to lower case. Lowercasing greatly improves output consistency and may be used for the majority of NLP as well as text mining activities.

3.5 Feature extraction using TF-IDF

The features are extracted from these pre-processed data using the TF-IDF approach. The process of extracting features starts with parsing every text and eliminating a group of pre-specified stop words from a semantic standpoint. Next, from the collection of retrieved features, representative features are chosen. To eliminate noisy features, feature selection describes a crucial pre-processing technique. It increases data comprehension and lowers the high dimensionality related to the feature space, which enhances the efficiency, performance, and outcome of clustering. It has been extensively employed in supervised learning tasks like text categorization. As a result, it remains crucial for raising the efficacy as well as efficiency of clustering. The feature selection metrics TF-IDF and its variants are often used.

The dataset to be clustered has been described as a collection of vectors, $Y = \{y_1, y_2, \cdots, y_o\}$, in major clustering methods, in which the vector y_j have been referred to as the feature vector associated with a single item. A vector e, like $e = \{x_1, x_2, \cdots, x_o\}$, represents the content of a document as a dot in a multidimensional space, in which x_j shows the word weight related to the term u_j. The word's frequency of occurrence both within a document and throughout the group of documents have been considered when calculating the term weight. The combination of the TF-IDF describes the major utilized weighting method. The term's frequency in a given document provides an indication of its significance. A statistical metric called TF-IDF illustrates a word's significance to a manuscript. Frequent words in a document are more significant because they are more representative of the subject.

Let g_{jk} be the word j's frequency in document k. Next, standardize term frequency (tf) throughout the corpus as a whole:

$$tf_{jk} = g_{jk} / \max\{g_{jk}\} \tag{1}$$

The term's overall relevance has been gauged by the inverse document frequency. Terms that are used throughout several publications don't always accurately reflect the subject matter. Assume df_j be the count of documents that include term j and document frequency of j. idf_j represent the term j's inverse document frequency.

$$idf_j = \log_2(0/df_j) \tag{2}$$

Here, the term 0 represent the total count of documents. TF-IDF weighting describes a common integrated term significance indicator.

$$x_{jk} = tf_{jk} \times idf_j = tf_{jk} \times \log_2(0/df_j) \tag{3}$$

3.6 Clustering using Spectral Clustering

The extracted features undergo the final clustering phase using the spectral clustering algorithm, in which the parameters of spectral clustering algorithm are tuned by PSO with the consideration of silhouette score maximization as the major fitness function, thus called to be novel spectral clustering-PSO. Spectral clustering is divided into three components: To begin, create the similarity graph X as well as the Laplacian matrix M =

$E - X$, where E describes the degree matrix. Next, calculate the Laplacian matrix's initial l eigenvectors. Furthermore, employ k-means to the matrix, which has been made up of the l eigenvectors. The loss function of spectral clustering may be expressed as follows:

$$M_d(\theta) = \frac{1}{n^2}\sum_{j,k=1}^{n} X_{j,k}\left\|z_j - z_k\right\|^2 \quad (4)$$

The output associated with the spectral clustering network G_θ is denoted by z_j, while n represents the minibatch size. The similarity among y_j and y_k is captured by the formula $X_{j,k} = \omega(y_j, y_k)$. Regarding ω, the objective is to translate related points y, y' (that is, points with big values of $\omega(y, y')$ into an embedding space in which those points are near to one another.

Similarly, instance y_j should be embedded near to its neighbors O_{pre}, O_{local}, and O_{global}, and its augmentation data. The loss function is seen on instance level, which has been derived from the similar objective and not only attained the similar goal yet also given more careful consideration. Moreover, an orthogonal restriction is imposed by spectral clustering to thwart simplistic solutions in which every information point is mapped to the similar output vector:

$$\frac{1}{n}Z^T Z = J_{l\times l} \quad (5)$$

Z describes an output matrix of size $n \times l$, in which z_j^T shows the j^{th} row. The network's last layer has been used to provide orthogonality enforcement. The last layer acts as a linear layer, receiving input from l units and producing l outputs. The weights are adjusted to get the orthogonalized findings Z for each batch. The inputs to this layer are represented by the $n \times l$ matrix $\tilde{Z}$. To ensure column-wise orthogonality of $\tilde{Z}$'s columns, a linear map has been preferred utilizing QR decomposition. Specifically, the QR decomposition of any matrix B satisfying B^TB having full rank may be obtained by applying the Cholesky decomposition:

$$B^TB = DD^T \quad (6)$$

When R is placed to $R = B(D^{-1})^T$, in which D describes a lower triangular matrix, Q may be derived. Thus, to orthogonalize matrix $\tilde{Z}$, the final layer does a right multiplication of $\sqrt{n}(\tilde{M}^{-1})^T$. The $\tilde{M}$ is obtained from the Cholesky breakdown of $\tilde{Z}$, and the $\sqrt{n}$ term has been included to satisfy Equation (5). The QR decomposition have been used to modify the final layer's weights at each orthogonalization stage. Entire weights, including the weights associated with the final layer, which functions just as a linear layer, are frozen once the Neural Network (NN) has finished training. Additionally, this layer fosters clustering assignments that are easier to discern.

Because spectral clustering decreases the dimensionality related to the information prior to clustering, it can handle huge datasets as well as high-dimensional information. Since it doesn't depend on conventional distance-oriented clustering techniques, it may be used with non-linearly separable information. Its slowness shows one of its drawbacks. It becomes necessary to use a speedier method if the dataset has a lot of data points. It is dependent on the selection of parameters, including the count of clusters as well as the kernel type, and ineffective parameter selection might lead to less-than-ideal clustering outcomes. Hence, to overcome the limitations of traditional spectral clustering algorithm, the parameters are tuned by PSO with the consideration of silhouette score maximization as the major fitness function, thereby called as novel spectral clustering-PSO. This innovative spectral clustering-PSO can handle vast datasets and also overcome the problem of computational complexity. This proposed spectral clustering-PSO clusters the final output into six classes such as data mining, deep learning, image, machine learning, network, and sports.

3.7 PSO algorithm

The PSO algorithm has been selected here in the proposed published document clustering model for tuning the parameters of the spectral clustering algorithm with the intention of silhouette score maximization as the major fitness function. The PSO approach was first designed to mimic the social characteristic related to a flock of birds. However, after simplifying the algorithm, it was discovered that the individual particles—referred to here as particles—were really carrying out optimization.

The PSO technique starts with the particles in the search space at random places and moves them in randomly determined paths. A particle's path is next progressively altered such that it begins to go towards its optimal past

locations as well as those of its peers. It searches the area around these locations in the hopes of finding even better locations in relation to certain fitness measure g: $S^o \rightarrow S$.

Assume $\vec{w}$ be a particle's velocity and consider $\vec{y} \in S^o$ represent its location. Both are first selected at random and are next updated repeatedly using two formulas. The particle's velocity may be updated using the below formula.

$$\vec{w} \leftarrow \omega\vec{w} + \phi_q s_q(\vec{q} - \vec{y}) + \phi_h s_h(\vec{h} - \vec{y}) \quad (7)$$

Here, the particle's velocity recurrence has been controlled by the user-defined behavioral parameter $\omega \in S$, also known as the inertia weight. The earlier optimal location of the particle is represented by $\vec{q}$, and the earlier optimal location of the swarm, via which the particles implicitly interact with one another, is $\vec{h}$. The user-defined behavioral parameters $\phi_q, \phi_h \in S$ and the stochastic variables $s_q, s_h \sim V(0,1)$ are used to weight these. Regardless of whether the particle's fit gets better or worse, including the velocity to its existing location moves it to a different location in the search space:

$$\vec{y} \leftarrow \vec{y} + \vec{w} \quad (8)$$

It becomes standard practice to enforce search-space bounds after updating a particle's location. In order to do this, a particle's velocity $\vec{w}$ is confined to the whole dynamic range associated with the search-space, allowing it to go across search space boundaries no more than once in a single step. Algorithm 1 displays the PSO pseudocode.

Algorithm 1: PSO

Start

Particle initialization using random velocities and positions [extracted features of the proposed published document clustering model]

Until a termination condition has been attained [maximized silhouette score of the proposed published document clustering model]

For every particle having velocity $\vec{w}$ and position $\vec{y}$ do

$$\vec{w} \leftarrow \omega\vec{w} + \phi_q s_q(\vec{q} - \vec{y}) + \phi_h s_h(\vec{h} - \vec{y})$$

Enforce the velocity boundaries

$$\vec{y} \leftarrow \vec{y} + \vec{w}$$

Enforce the search-space boundaries

If $(g(\vec{y}) < g(\vec{q}))$ then update the optimal position of the particle [maximized silhouette score of the proposed published document clustering model]

$$\vec{q} \leftarrow \vec{y}$$

And similarly for the complete swarm's optimal position $\vec{h}$ [silhouette score maximization of the proposed published document clustering model]

End for

Stop

4. RESULT ANALYSIS

4.1 Experimental setup

The proposed spectral clustering-PSO model was implemented in Python and the findings were analyzed. The population size and the iteration count was placed to 10 and 100. The total count of clusters considered represents the count of classes i.e., 6. The proposed spectral clustering-PSO model was compared with various traditional methods like JA-GWO [11], tpLDA [12], HDMA [14], and Net2Vec [16] in terms of distinct analysis

such as silhouette score, davies bouldin score, homogeneity score, and completeness score to reveal the betterment of the suggested model.

4.2 Silhouette score analysis

The silhouette score analysis of the proposed spectral clustering-PSO and the existing models for the published document clustering is shown in Fig. 3. The distance separating the produced clusters may be validated employing silhouette analysis. It provides a visual means of evaluating factors like the count of clusters by describing a metric associated with the proximity within every point in a cluster as well as points in the neighboring clusters. The range related to this measure is [-1, 1]. When the silhouette coefficient is near to +1, it shows that the sample is not close to any close clusters. Negative values imply that the samples may have been inaccurately allocated to one among the neighboring clusters, wherein values of 0 imply that the sample is on or very on the decision border among two neighboring clusters. The proposed spectral clustering-PSO in terms of silhouette score has been 51.92%, 70.81%, 45.93%, and 20.89% better than JA-GWO, tpLDA, HDMA, and Net2Vec respectively. Thus, it can be concluded that, the proposed spectral clustering-PSO model outperforms the other methods in terms of silhouette score for the developed published document clustering model.

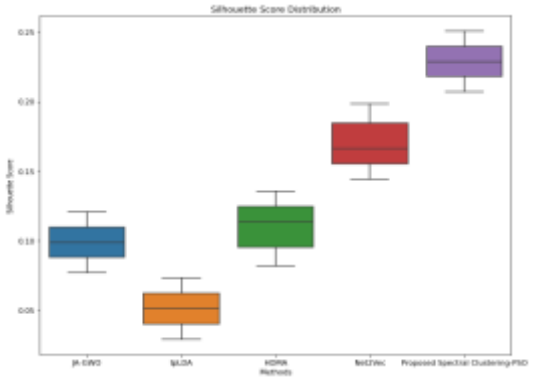


Fig. 3. Silhouette score analysis

4.3 Davies bouldin score analysis

Fig. 4 displays the results of the davies bouldin score study comparing the traditional models for the published document clustering with the suggested spectral clustering-PSO. It has been described by comparing every cluster's average similarity metric to that of its close identical cluster. The ratio of within-cluster to between-cluster distances is employed to describe similarity. Consequently, clusters with greater distances and reduced dispersion will score more. A lower score corresponds to better clustering, with zero being the least. The proposed spectral clustering-PSO in terms of davies bouldin score is 89.69%, 58.48%, 32.67%, and 13.99% advanced than JA-GWO, tpLDA, HDMA, and Net2Vec respectively. Therefore, it can be said that, for the created published document clustering model, the suggested spectral clustering-PSO model performs better than the other approaches in terms of davies bouldin score.

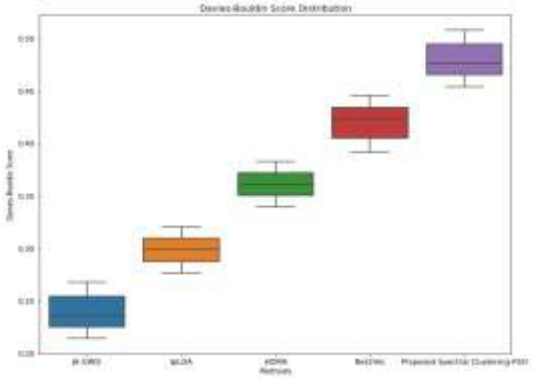


Fig. 4. Davies Bouldin score analysis

4.4 Homogeneity score analysis

The results of the homogeneity score investigation comparing the recommended spectral clustering-PSO with the conventional models for the published document clustering are shown in Table 1. When all of a clustering result's clusters include only data points from one class, the result fulfils homogeneity. A permutation associated with the class or cluster label values has no effect on the score value, meaning that this metric is not dependent of the absolute values associated with the labels. There exists no symmetry in this measure. The proposed spectral clustering-PSO with respect to homogeneity score is 79.21%, 43.20%, 19.87%, and 2.61% higher than JA-GWO, tpLDA, HDMA, and Net2Vec respectively. Thus, it can be concluded that the proposed spectral clustering-PSO model outperforms the other methods with respect to homogeneity score for the developed published document clustering model.

Table 1 Homogeneity score analysis

Methods	**Clusters**					
	1	**2**	**3**	**4**	**5**	**6**
JA-GWO [11]	0.0070	0.0157	0.0244	0.0331	0.0418	0.0505
tpLDA [12]	0.0192	0.0280	0.0368	0.0456	0.0544	0.0632
HDMA [14]	0.0220	0.0307	0.0494	0.0581	0.0668	0.0755
Net2Vec [16]	0.0342	0.0430	0.0518	0.0606	0.0794	0.0882
Proposed Spectral Clustering-PSO	0.0470	0.0557	0.0644	0.0731	0.0818	0.0905

4.5 Completeness Score analysis

Table 2 displays the findings related to the completeness score analysis contrasting the suggested spectral clustering-PSO with the traditional models for the published document clustering. If every data point that belongs to a class represents an element associated with the identical cluster, then the clustering result seems to be complete. There exists no symmetry in this metric. The proposed spectral clustering-PSO in terms of completeness score is 62.21%, 45.94%, 32.36%, and 16.09% superior to JA-GWO, tpLDA, HDMA, and Net2Vec respectively. Therefore, it can be said that, in terms of completeness score for the established published document clustering model, the suggested spectral clustering-PSO model accomplishes better than the remaining approaches.

Table 2 Completeness score analysis

Methods	**Clusters**					
	1	**2**	**3**	**4**	**5**	**6**
JA-GWO [11]	0.0921	0.1008	0.1195	0.1282	0.1369	0.1456
tpLDA [12]	0.1543	0.1631	0.1719	0.1807	0.1995	0.2083
HDMA [14]	0.2171	0.2258	0.2345	0.2432	0.2519	0.2606
Net2Vec [16]	0.2793	0.2881	0.2969	0.3057	0.3145	0.3233
Proposed Spectral Clustering-PSO	0.3321	0.3408	0.3595	0.3682	0.3769	0.3853

5. CONCLUSION

In this research paper, an optimization approach inspired by nature was used to conduct document clustering. The dataset was first collected manually from various sources. The next step involved preparing the data so that the text content of the published documents could be extracted. Pre-processing was done on this prepared data to eliminate stop words, punctuations, and lowercase conversion. The TF-IDF approach was utilized to extract features from the pre-processed data while keeping the keywords intact. The collected features

proceeded through the last stage of clustering using the spectral clustering method. The objective function of this phase was silhouette score maximization, and the parameters were tuned using the PSO algorithm, a nature-inspired optimization technique. The final result was clustered into six groups using the suggested spectral clustering-PSO, including data mining, deep learning, image, machine learning, network, and sports. The advantage of the proposed document clustering model over the other techniques was explained in terms of unique metrics. The proposed spectral clustering-PSO in terms of silhouette score was 51.92%, 70.81%, 45.93%, and 20.89% better than JA-GWO, tpLDA, HDMA, and Net2Vec respectively. Similarly, the proposed spectral clustering-PSO in terms of davies bouldin score was 89.69%, 58.48%, 32.67%, and 13.99% advanced than JA-GWO, tpLDA, HDMA, and Net2Vec respectively.

REFERENCES

[1] Abualigah, L.M., Khader, A.T. and Hanandeh, E.S. "A combination of objective functions and hybrid Krill herd algorithm for text document clustering analysis", Eng. Appl. Artif. Intel., vol. 73, pp. 111-125, 2018.

[2] Brockmeier, A.J., Mu, T., Ananiadou, S. and Goulermas, J.Y. "Self-tuned descriptive document clustering using a predictive network", IEEE Trans. Knowl. Data Eng., vol. 30, pp. 1929-1942, 2018.

[3] S. Mayani and S. Swarndeep, "A survey of text document clustering by using clustering techniques", Int. Res. J. Eng. Technol., vol. 6, no. 12, pp. 1-5, Dec. 2019.

[4] R. Lakshmi and S. Baskar, "DIC-DOC-K-means: Dissimilarity-based initial centroid selection for document clustering using K-means for improving the effectiveness of text document clustering", J. Inf. Sci., vol. 45, no. 6, pp. 818-832, Dec. 2019.

[5] L. Abualigah, A. H. Gandomi, M. A. Elaziz, A. G. Hussien, A. M. Khasawneh, M. Alshinwan, and E. H. Houssein, "Nature-inspired optimization algorithms for text document clustering-A comprehensive analysis", Algorithms, vol. 13, no. 12, p. 345, Dec. 2020.

[6] D. S. Rajput, "Review on recent developments in frequent itemset based document clustering, its research trends and applications", Int. J. Data Anal. Techn. Strategies, vol. 11, no. 2, pp. 176-195, 2019.

[7] T. Sutanto and R. Nayak, "Fine-grained document clustering via ranking and its application to social media analytics", Social Netw. Anal. Mining, vol. 8, p. 29, Dec. 2018.

[8] E. L. Lydia, P. K. Kumar, K. Shankar, S. K. Lakshmanaprabu, R. M. Vidhyavathi, and A. Maseleno, "Charismatic document clustering through novel K-means non-negative matrix factorization (KNMF) algorithm using key phrase extraction", Int. J. Parallel Program., pp. 1-19, Aug. 2018.

[9] Manoharan, J. S. "Double attribute-based node deployment in wireless sensor networks using novel weight-based clustering approach", Sadhana, vol. 47, no. 3, pp. 1-11, 2022.

[10] Devi, C. U., Manoharan, S., Thilagam, V. P. “A novel optimized cluster-based trust model in a wireless sensor network with GSTEB routing protocol”, Journal of advanced research in dynamical and control systems, vol. 11, no. 4, pp. 1245-1255, 2019.

[11] G. Venkanna and D. K. F. Bharati, "Optimal Text Document Clustering Enabled by Weighed Similarity Oriented Jaya with Grey Wolf Optimization Algorithm," in The Computer Journal, vol. 64, no. 1, pp. 960-972, Nov. 2019.

[12] P. Yang, Y. Yao and H. Zhou, "Leveraging Global and Local Topic Popularities for LDA-Based Document Clustering," in IEEE Access, vol. 8, pp. 24734-24745, 2020.

[13] F. Malik, S. Khan, A. Rizwan, G. Atteia and N. A. Samee, "A Novel Hybrid Clustering Approach Based on Black Hole Algorithm for Document Clustering," in IEEE Access, vol. 10, pp. 97310-97326, 2022.

[14] R. Huang, W. Xu, Y. Qin and Y. Chen, "Hierarchical Dirichlet Multinomial Allocation Model for Multi-Source Document Clustering," in IEEE Access, vol. 8, pp. 109917-109927, 2020.

[15] A. M. Sheri, M. A. Rafique, M. T. Hassan, K. N. Junejo and M. Jeon, "Boosting Discrimination Information Based Document Clustering Using Consensus and Classification," in IEEE Access, vol. 7, pp. 78954-78962, 2019.

[16] Y. C. Yoon, H. K. Gee and H. Lim, "Network-Based Document Clustering Using External Ranking Loss for Network Embedding," in IEEE Access, vol. 7, pp. 155412-155423, 2019.

[17] Muruganantham Ponnusamy, Pradeep Bedi, Tamilarasi Suresh, Aravindhan Alagarsamy, R. Manikandan & N. Yuvaraj,"Design and analysis of text document clustering using salp swarm algorithm", The Journal of Supercomputing, vol. 78, pp. 16197-16213, May 2022.

[18] Meena Chaudhary, Jyoti Pruthi, Vinay Kumar Jain & Suryakant, "A novel squirrel search clustering algorithm for text document clustering", International Journal of Information Technology, vol. 14, pp. 3277-3286, August 2022.

[19] Lina Ren, Yongbin Qin, Yanping Chen, Ruina Bai, Jingjing Xue & Ruizhang Huang, "Deep structural enhanced network for document clustering", Applied Intelligence, vol. 53, pp. 12163-12178, September 2022.

[20] Seyed Mojtaba Sadjadi, Hoda Mashayekhi & Hamid Hassanpour, "A semi-supervised framework for concept-based hierarchical document clustering", World Wide Web, vol. 26, pp. 3861-3890, October 2023.