Reviews on swarm intelligence algorithms for text document clustering

S DHANALAKSHMI[1], S SATHIYABAMA[2]

[1,2] Department of Computer Science
[1,2] Thiruvalluvar Government Arts College, Rasipuram, Tamilnadu, India

---

Abstract - *Text clustering is an unsupervised learning technique that divides a large number of text documents into a small number of clusters. Each cluster contains similar documents, while the clusters contain dissimilar text documents. Various optimization problems, including text document clustering challenges, have been effectively solved using swarm intelligence (SI) optimization algorithms. This paper reviews all of the relevant literature on SI-based text document clustering applications including many variants, basic, enhanced, and hybrid methods. The main procedure of text clustering, distance and similarity function, and theoretical discussion are also discussed.*

Key Words: text mining, text clustering, swarm intelligence, optimization algorithms, data mining

1. INTRODUCTION

Clustering is a general text mining technique for representing a dataset using a limited set of clusters, sometimes with a fixed number of clusters, based on similarities between its elements [1, 2]. The partitioning clustering technique is widely applied to solve real-world applications including data clustering, image clustering, marketing, and bio-informatics. The goal of text clustering is to create optimal clusters that contain related documents. Clustering is the process of dividing a large number of documents into a set of related groups.

Each group contains many similar objects, but different groups contain different objects [3-5]. The overall technique for the text clustering problem as an optimization problem, its formulation, mathematical notations, preprocessing stages, document format, clustering problem solution representation, and the fitness function are all described in this part. This can aid future studies in obtaining clear broad information about the issue [6, 7].

Extracting meaningful data from documents is a difficult operation that necessitates the use of rapid and high-quality document clustering techniques. The K-means algorithm is a straightforward, quick, and unsupervised partitioning algorithm that produces results that are both parallelized and comprehensible [8, 9]. However, it has many drawbacks like local optima, low accuracy, and failure to achieve global solutions. Hence, the swarm intelligence (SI) optimization algorithm can overcome the limitations of the k-means algorithm and obtain a global optimum value.

SI is the collective behavior of self-organized and decentralized systems, which includes both intelligent and non-intelligent individuals who follow simple rules or behaviors to do very complicated tasks with limited local information. Observing natural or artificial behaviors such as bird flocks, fish schools, and ant food foraging led to the development of SI algorithms. Particle swarm optimization (PSO), bat optimization (BA), grey wolf optimization (GWO), firefly optimization (FFA), ant colony optimization (ACO), artificial fish swarm algorithm (AFSA), and artificial bee colony optimization (ABC) are examples of SI algorithms [10].

By comparing the many SI optimization algorithms accessible for text document clustering, this research aims to give the reader an accurate overview of the numerous SI optimization algorithms available. This work investigates the accessibility of each class and the implementation of a suitable optimization algorithm for each. As a result, this research will help academics and clinicians choose methods and algorithms that are appropriate for a wide range of text clustering applications. As a result of these studies, the goal of this study was to investigate the field of SI clustering methods and achieve the following goals:

- The present paper discusses the pre-processing steps of text clustering
- Various related works have been discussed using swarm intelligence algorithms.
- To provide a comprehensive classification of clustering evaluation criteria that can be used in experimental research.
- Conduct a theoretical examination of each class's best representative SI optimization techniques.

The following is a list of the main sections of this study. In Section 2, the main procedures of the text clustering are presented. The variants of SI algorithms that have been employed to solve text clustering problems are shown in Section 3. Section 4 discusses the evaluation criteria utilized in text clustering applications. Section 5 contains a discussion and theoretical analysis. Finally, Section 6 presents the survey's findings as well as future research possibilities.

2. PROBLEME FORMULATION

A collection of documents D is grouped into a predefined number of K clusters. D can be demonstrated as an object's vectors $D = \{d_1, d_2, ..., d_n\}$. Here, n is the total number of documents. Each group has a cluster centroid, which is denoted by c_k and is represented as a vector of word term weights. c_k presents the K^{th} cluster center. The similarity of documents is determined using distance measures, which are used to cluster each object to the cluster centroid that is closest to it. The fundamental goal of clustering is to create groups based on the intrinsic contents of the objects. The text requires model pre-processing actions for constructing clusters in this era. Tokenizing, stop word removing, stemming, item weighting and document representation are pre-processing steps in text clustering. The pre-processing procedures are briefly discussed in the next sub-section.

2.1 Tokenizing

Tokenization is the method of breaking words into bits (or tokens) that are probably missing individual letters at the same time, such as punctuation. These tokens are usually linked to words or items, but it's important to distinguish between types/tokens. A token is a sequence of letters that is structured as a functional semantic unit in a text. A sort is a collection of all tokens that have the same letter chain. A word is an example of a term found in the search method's vocabulary.

2.2 Stop words removal

Stop words include terms like “which”, “the”,”our”, “is”, "an,", “me”, "that," and others, as well as important phrases in the text that are extremely often used and small beneficial words. These terms should be removed from the text since they are frequently repeated, reducing the effectiveness of clustering approaches. The total number of terms on the stop-word list is above 500.

2.3 Stemming

Stemming is the procedure of shortening current terms to their root/stem. The stem technique differs from the root morphological method in that it uses the same stem to outline words, although it is not a true root.

2.4 Term weighting

The frequency of each phrase in the document is used to provide a term weighting to each term or feature. In weighting procedures, the term frequency-inverse document frequency (TF-IDF) is commonly used. The following equation shows how each document is represented as a vector of term weights.

$$d_i = (w_{i,1}, w_{i,2}, ..., w_{i,t}) \qquad (1)$$

The term weight in the document i for the feature j is calculated as follows,

$$w_{i,j} = tf(i,j) \times idf(i,j) = ft(i,j) \times \log(\frac{n}{df(j)}) \qquad (2)$$

Where, $w_{i,j}$ -denotes the weight of document i and term j. $tf(i,j)$ -denotes the occurrence of term j in document i. $idf(i,j)$ - denotes the inverse document frequency. n represents the number of all documents in the datasets. $df(j)$ -represents the number of documents that contain features j. In most text mining applications, the vector space model (VSM) is utilized to describe document features as a vector (row) of weights. The following equation is the VSM's most frequent format, which consists of n documents with t words in each.

$$VSM = \begin{bmatrix} \omega_{1,1} & \omega_{1,2} & \cdots & \omega_{1,(t-1)} & \omega_{1,t} \\ \vdots & \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \ddots & \cdots & \cdots \\ \omega_{(n-1),1} & \omega_{(n-1),2} & \cdots & \cdots & \omega_{(n-1),t} \\ \omega_{n,1} & \omega_{n,2} & \cdots & \omega_{n,(t-1)} & \omega_{n,t} \end{bmatrix} \qquad (3)$$

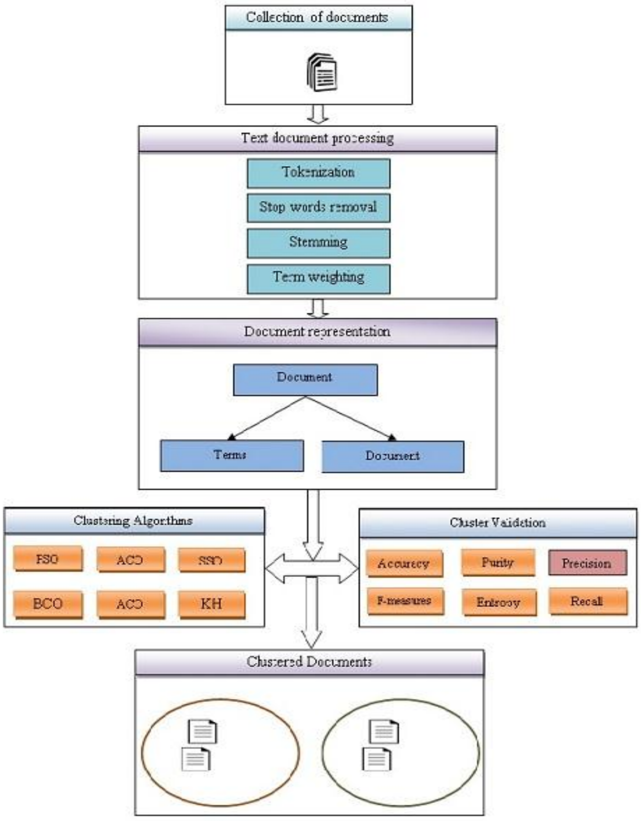

Figure 1 : Overviews of text document clustering

3. RELATED WORKS FOR TEXT CLUSTERING PROBLEME

Text document clustering is a method of separating documents into parent groups, may manage documents with unlabeled or unauthorized clusters, which is unusual. The determination of the similarity measure between the documents in each cluster has an impact on the text clustering process. Text clustering is a technique for improving the consistency of text document organization and providing a user-friendly view of documents based on an evaluation criterion that can be represented as an objective function. Text clustering-based methods, in particular, have aided users in dealing with a huge number of documents and organizing them with better precision and simplicity than if they did not utilize a text clustering-based model. Because of its minimal computational cost, the partitional clustering approach is now well-suited for clustering a huge collection of document data sets [50]. Furthermore, because most partitional clustering techniques are essentially linear in time, they are commonly used. Figure 1 shows the overviews of text document clustering process.

The PSO is a well-known SI method that is applied to solve various real-world applications including text document clustering [11]. For example, Karol et al. (2013) [12] developed a new method which is the use of a soft computing strategy as an intelligent hybrid approach. The hybrid approach combined with PSO and K-Means and FCM (fuzzy c-means). Cagnina, L et al. (2014) [13] developed a new efficient text clustering method based on PSO (CLUDIPSO) for short text news. With small corpora, CLUDIPSO produced high-quality results, but with bigger corpora, performance deteriorated significantly. Song et al. (2015) developed a new hybrid text \clustering algorithm based on Quantum-behaved PSO (QPSO) and genetic algorithm (GA) [14]. T[he created method is being evaluated to normalize the search range of particles in a sound environment using a new position updating strategy. Such an approach can increase the likpoelihood of finding the best option. Pamba et al. (2017) [15] developed a document ring method using a frequent pattern growth strategy based on fuzzy PSO (FPFPSO). The goal is to assess the proposed methodology for clustering online documents with varying degrees of membership in all of their respective clusters. The approach reduces search space by preserving relationships with the search context and automatically generates cluster centroids and particles.

Chouhan et al. (2018) [16] introduces a method for document clustering based on the PSO and k-means. The PSO method is used before K-means to locate the best points in the search space, and these points are utilized as initial cluster centroids in the K-means to find final document clusters. Janani et al. (2019) [17] introduce a novel spectral clustering algorithm with PSO (SCPSO) to improve text document clustering. The randomization is done using the initial population using global and local optimization functions. To deal with the massive amount of text documents, the goal is to combine spectral clustering with swarm optimization. The final findings reveal that the suggested SCPSO algorithm outperforms other clustering algorithms in terms of clustering accuracy. Pamba et al. (2020) [18] developed a new document clustering method using Self-adaptive Frequent Pattern Growth-Based Dynamic Fuzzy PSO (FPDFPSO) for Web Document Clustering. The FPDFPSO outperforms existing conventional approaches in terms of convergence speed and diversity preservation.

Alam et al. (2016) [19], a web clustering technique, is presented to cluster web content effectively. Both K-means and the Artificial Bee Colony (ABC) clustering algorithms are used in the suggested technique. The ABC method is utilized as the global search optimizer in this paper, while K-means is used to refine the solutions. As a result, the cluster's quality has improved. Hijazi et al. (2021) [20] introduces a feature selection method based on the combination of chi-square and an artificial bee colony (ABC). The results showed that a smaller number of features surpassed the original features set in terms of classification accuracy. Furthermore, when compared to the chi-square approach and the ABC algorithm as a feature selection method, the proposed method performed better. Mohammed et a. (2015) developed a new text clustering method based weight-based firefly algorithm (WFA) [21]. This finding method achieved better clustering by search solution's exploitation is improved. Mohammed et al. (2016) [22] introduced an improved search method of the ABC and embed two local search paradigms namely gradient and chaotic local search to improve its exploitation capability.

The SSO algorithm is based on a simulation of social-spider cooperative behaviour. Individuals in the proposed algorithm imitate a group of spiders that interact with one another depending on the cooperative colony's biological laws [23]. Chandran et al. (2017) [24] developed a novel text clustering algorithm based on SSO. Chandran, T.R. et al. (2018) developed a new document clustering method based on [25] SSO. The use of a single cluster implementation for each spider is defined in this study as a new and more effective implementation of SSO (ESSOSC) for solving the text document clustering problem. Rashaideh et al. (2020) developed a new text clustering algorithm-based grey wolf optimizer (GWO) [26]. The GWO is a successful and effective swarm-based algorithm that was recently presented to simulate the hunting behaviour of grey wolves. GWO has enhanced the allocation of documents to the best clusters, as well as assisted in the transition from local optimum solutions to the best global solutions.

Krill herd (KH) is a revolutionary swarm-based optimization system that mimics the behaviour of a KH in real-time. This algorithm has a lot of potential because it outperforms other optimization methods and balances the exploration and exploitation process by combining the strengths of local nearby searching and global wide-range searching. The KH has been applied to solve the document clustering problem. For example, Abualigah et al. (2016) developed a new innovative text clustering method based on the KH to improve the clustering of web text documents [27]. The basic KH algorithm with all of its operators is used in the first technique, while the genetic operators in the basic KH algorithm are ignored in a second way. The proposed KH algorithms' performance is evaluated and compared to the k-mean algorithm.

Abualigah, L.M et al. (2018) [28] developed a new hybrid text clustering method based new objective function and KH algorithm. In comparison to the other comparative algorithms, the suggested modified KH with hybrid function acquired almost all of the best outcomes for all datasets. More, P.S et al. (2021) [29] developed a new text clustering method based on the KH. Pre-processing is used to remove noise and artifacts from the data, which is done using the stemming and stop word removal methods. Furthermore, the term frequency-inverse document frequency (TF–IDF) method is used to extract notable features. TF–IDF features are used to cluster documents, while the KH approach is used to cluster text documents. The fitness function was created from scratch, taking into account mean square error and Jaccard similarity.

Feature selection is a key unsupervised learning strategy for selecting a fresh subset of useful text characteristics to improve text clustering performance and reduce computational time. To improve the performance and computing efficiency of text clustering algorithms, a variety of feature selection methods are used. For example, Abualigah et al. (2017) [30] developed a new feature selection method using a combination of PSO and genetic algorithm for enhancing the performance of text clustering problems. Asif, M et al. (2021) [31] introduces a self-inertia weight adaptive PSO (SIW-APSO) based feature selection approach to improve the performance of text classification systems. Because of its high search competency and ability to efficiently locate feature subsets, SIW-APSO has a fast convergence phenomenon.

Singh et al. (2021) [32] presents a new hybrid feature selection technique by utilizing the combination of ACO and GA. The feature selection method has been improved, resulting in a reduction in the dimensionality of the feature space and, as a result, an increase in performance. Janani et al. (2020) [33] introduced a new automatic text classification approach based on machine learning. There are two stages to the field study. The relevant features for classification are chosen in the first phase, and text documents are classified in the second phase.

4. EVALUATION METRICS

The text clustering technique offers two types of evaluation metrics: internal and exterior metrics. Details of these measures are discussed in the following sections,

4.1 Internal metrics

The underlying base for developing the clustering optimization algorithms processes is similarity and distance metrics. When it comes to qualitative results, the distance measure is preferred when dealing with qualitative data. Internal metrics are employed to determine a group of text documents to be delivered to their typical cluster without regard to the text class mark.

4.1.1 Distance measures

Clustering necessitates the construction of a distance measure, which lends a numerical value to the extent of difference between two documents and is used by the clustering algorithm to divide a dataset into groups. There is no single distance measure that works well in all situations, therefore choosing one is a matter of determining which distance measure best captures the essence of essential distinguishing traits for the particular document set. Table 1 shows some mostly used distance measures in literature.

4.1.2 Similarity Measures

In-text mining, determining the degree of similarity between two documents is a difficult task. Mostly used similarity measures are reviewed in collected articles and demonstrated in Table 2 [34, 35].

4.2 External measures

External measurements are used to examine the accuracy (correctness) of the gathered clusters based on the class labels in the dataset for the current article. The external evaluation measures used to examine the output of the clustering algorithms are described in the subsections below. Accuracy, purity, entropy, precision, recall, and F-measures are the most commonly used evaluation metrics in the text clustering domain.

4.2.1 Accuracy

To determine the proper documents assigned to all categories in the given dataset, accuracy is used. This accuracy metric is calculated as follows:

$$AC = \frac{1}{n}\sum_{i=1}^{k} n_{i,i} \tag{4}$$

Where, $n_{i,i}$ is the total number of correct candidates for class i in cluster. n is the total number of given documents, and K is the total number of given clusters in the dataset.

4.2.2 Purity

In a large class, the purity is used to compute each cluster's percentage. This metric assigns each group to the most common classification. Because of the percentage of large class sizes in each group, which is compared according to its size, an acceptable purity value is close to 1. As a result, the purity value in the $\left[\frac{1}{k^{+}},1\right]$ interval is used to derive the purity value of the cluster j:

$$P(c_j) = \frac{1}{n_j}\max_{j} n_{i,j} \tag{5}$$

Where, $\max_j$ - denotes the size of the large class in group j. n_{ij} - denotes the number of all exact of the class label i in cluster j. n_{ij} - denotes the complete number of members of cluster j.

Table 1: Distance measurements

Name	Formula
Minkowski distance	$=\left(\sum_{l=1}^{d}\left\|x_{li}-x_{lj}\right\|^{n})^{1/n}\right)$
Standardized Euclidean distance	$=\left(\sum_{l=1}^{d}\left\|\frac{x_{li}-x_{lj}}{sl}\right\|^{2})^{1/2}\right)$
Cosine distance	$=1-\cos\alpha \quad = \frac{x_i^T \quad x_j}{\|x_i\| \quad \|x_j\|}$
Pearson correlation distance	$=1 \quad - \frac{cov(x_i,x_j)}{\sqrt{D(x_i)} \quad \sqrt{D(x_j)}}$
Mahalanobis distance	$=\sqrt{(x_i-x_j)^T S^{-1}(x_i-x_j)}$

Table 2: Similarity measures

Name	Formula
Cosine similarity	$\cos(A,B)=\frac{A.B}{\|A\|\times\|B\|}$
Jaccard similarity	$J(A,B)=\left\|\frac{A\cap B}{A\cup B}\right\|$
Sorensen similarity	$S(A,B)=\frac{2\times\|A\cap B\|}{\|A\|+\|B\|}$
Dice similarity	$Dice(A,B)=2\times\frac{\|A\cap B\|}{\|A+\|B\|\|}$

The purity for all clusters is determined as follows,

$$p=\sum_{j=1}^{k}\frac{n_j}{n}p(c_j) \tag{6}$$

4.2.3 Entropy

In each group, the entropy measures the partitioning of class labels. This metric focuses on how well different cluster classes are contained. A good sample has zero entropy, indicating a low entropy state for an optimal document clustering solution. The following equation can be used to calculate the entropy rate of a cluster j based on the size of each group.

$$E(c_j)=-\sum_i p_{i,j}\log p_{i,j} \tag{7}$$

Where p_{ij} is the probability value of class i members that belong to group j. The entropy for all groups is determined as follows,

$$E=\sum_{j=1}^{K}\frac{n_j}{n}E(c_j) \tag{8}$$

4.2.4 Precision

For each group, the precision is determined based on the specified class label in the main datasets. In all groups, the precision test is the ratio of documents to the total number of documents. The precision is determined for class i and class j as follows,

$$p(i,j)=\frac{n_{ij}}{n_j} \tag{9}$$

Where, n_{ij} -denotes the number of correct of the class i in the j group. n_j -denotes the total number of objects in the j group.

4.2.5 Recall

Based on the assigned class label, the recall (R) value for each cluster is calculated. The recall value is calculated by dividing the total number of relevant items in the dataset by the number of key documents in each category. The following equation is used to calculate the recall value for class i and cluster j.

$$p(i,j)=\frac{n_{ij}}{n_i} \tag{10}$$

Where, n_{ij} -denotes number of correct of the class i in the j group. n_j -denotes total number of objects in the j group.

4.2.6 F-Measures

The F-measures aim to evaluate clusters of the analyzed partition clusters at the class label partition clusters with the largest match. Based on the aggregation of precision and recall, this metric is a prominent evaluation criterion in the cluster domain. The following equation is calculated for f-measures:

$$F(j)=\frac{2\times P(i,j)\times R(i,j)}{P(i,j)+R(i,j)} \tag{11}$$

Where, $P(i, j)$ is precision value of class i and group j. $R(i, j)$ is recall value of class i and group j. Therefore, the f-measure for all clusters is determined as follows,

$$F = \frac{\sum_{j=1}^{K} F_j}{K} \qquad (12)$$

5. THEORETICAL DISCUSSIONS

Clustering analysis is a crucial technique for analyzing unlabeled data by creating a collection of clusters based on a predetermined number of clusters. Organizing pre-processing and algorithm advancement, as well as solution efficacy and evaluation, are all steps in the clustering process. Each is inextricably linked to the others and employs severe challenges in scientific research. SI algorithms developed by diverse research communities aim to tackle various clustering processes and have advantages and disadvantages. Because there are so many intrinsic conceivable conditions, there are still a lot of unsolved issues. These issues have already attracted and will continue to attract extensive applications from a wide range of fields. The current review and survey are organized by a list of key concerns and possible research topics for optimization cluster algorithms.

- New mechanisms have resulted in more sophisticated and difficult tasks, necessitating the use of more reliable clustering methods.
- No generally applicable clustering algorithm can address all clustering difficulties.
- Local optima, on the other hand, can get imprisoned due to their emphasis on exploration (i.e., global search) rather than exploitation (i.e., local search). This issue may improve over time if the sets of rules that govern the operation of various search algorithms become better understood. The initial cluster centroids and the number of clusters are two major issues in the text clustering application.
- The values and settings of the parameters affect the algorithm's overall performance, therefore parameter adjustment will be important in future investigations.
- At the post-processing and pre-processing stages, feature selection, extraction, and cluster validation are as important as the clustering techniques. The difficulty of succeeding schemes can be greatly reduced by selecting relevant and necessary qualities, and result evaluations reflect the level of trust we can place in the produced clusters. Regrettably, neither strategy has widespread leadership.
- Finally, the applications themselves are still responsible for balancing numerous criteria and processes.

6. CONCLUSIONS

This survey report examined over 25 research papers to determine the resilience and weaknesses of SI optimization techniques. This document comprehensively analyses the whole literature till the year 2021. The majority of the papers in this collection discuss optimization techniques based on SI that have been applied in text clustering (document) applications. Several algorithm variants are investigated, including usual, fundamental, enhanced, and hybrid. Furthermore, various new SI optimization techniques that can be used to solve clustering problems have recently been presented. To solve text clustering challenges, new hybrid and modified methods might be presented. Furthermore, various new meta-heuristic optimization techniques that can be used to solve clustering problems have recently been presented.

REFERENCES

[1] K. Tamilarisi, M. Gogulkumar, and K. Velusamy, "Data clustering using bacterial colony optimization with particle swarm optimization," in 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2021: IEEE, pp. 1-5.

[2] K. Tamilarasi, M. Gogulkumar, and K. Velusamy, "Enhancing the performance of social spider optimization with neighbourhood attraction algorithm," in Journal of Physics: Conference Series, 2021, vol. 1767, no. 1: IOP Publishing, p. 012017.

[3] D. Mustafi, A. Mustafi, and G. Sahoo, "A novel approach to text clustering using genetic algorithm based on the nearest neighbour heuristic," International Journal of Computers and Applications, vol. 44, no. 3, pp. 291-303, 2022.

[4] K. Velusamy and R. Manavalan, "Performance analysis of unsupervised classification based on optimization," International Journal of Computer Applications, vol. 975, p. 8887, 2012.

[5] S. S. Babu and K. Jayasudha, "A survey of nature-inspired algorithm for partitional data clustering," in Journal of Physics: Conference Series, 2020, vol. 1706, no. 1: IOP Publishing, p. 012163.

[6] A. A. Jalal and B. H. Ali, "Text documents clustering using data mining techniques," International Journal of Electrical & Computer Engineering (2088-8708), vol. 11, no. 1, 2021.

[7] S. Dhanalakshmi and S. Sathiyabama, "Survey on Information Mining Procedures Utilized in Healthcare Services."

[8] T.-H. Jo, "Inverted index based modified version of k-means algorithm for text clustering," Journal of Information Processing Systems, vol. 4, no. 2, pp. 67-76, 2008.

[9] K. Vijayakumari and V. Baby Deepa, "Fuzzy C-Means Hybrid with Fuzzy Bacterial Colony Optimization," in Advances in Electrical and Computer Technologies: Springer, 2021, pp. 75-87.

[10] S. Selvaraj and E. Choi, "Swarm Intelligence Algorithms in Text Document Clustering with Various Benchmarks," Sensors, vol. 21, no. 9, p. 3196, 2021.

[11] X. Cui, T. E. Potok, and P. Palathingal, "Document clustering using particle swarm optimization," in Proceedings 2005 IEEE Swarm Intelligence Symposium, 2005. SIS 2005., 2005: IEEE, pp. 185-191.

[12] S. Karol and V. Mangat, "Evaluation of text document clustering approach based on particle swarm optimization," Open Computer Science, vol. 3, no. 2, pp. 69-90, 2013.

[13] L. Cagnina, M. Errecalde, D. Ingaramo, and P. Rosso, "An efficient particle swarm optimization approach to cluster short texts," Information Sciences, vol. 265, pp. 36-49, 2014.

[14] W. Song, Y. Qiao, S. C. Park, and X. Qian, "A hybrid evolutionary computation approach with its application for optimizing text document clustering," Expert Systems with Applications, vol. 42, no. 5, pp. 2517-2524, 2015.

[15] R. V. Pamba, E. Sherly, and K. Mohan, "Evaluation of Frequent Pattern Growth Based Fuzzy Particle Swarm Optimization Approach for Web Document Clustering," Cham, 2017: Springer International Publishing, in Computational Science and Its Applications – ICCSA 2017, pp. 372-384.

[16] R. Chouhan and A. Purohit, "An approach for document clustering using PSO and K-means algorithm," in 2018 2nd International Conference on Inventive Systems and Control (ICISC), 2018: IEEE, pp. 1380-1384.

[17] R. Janani and S. Vijayarani, "Text document clustering using spectral clustering algorithm with particle swarm optimization," Expert Systems with Applications, vol. 134, pp. 192-200, 2019.

[18] R. V. Pamba, E. Sherly, and K. Mohan, "Self-adaptive Frequent Pattern Growth-Based Dynamic Fuzzy Particle Swarm Optimization for Web Document Clustering," in Computational Intelligence: Theories, Applications and Future Directions-Volume II: Springer, 2019, pp. 15-25.

[19] M. Alam and S. Baulkani, "A hybrid approach for web document clustering using K-means and artificial bee colony algorithm," Int J Intell Eng Syst, vol. 9, no. 4, pp. 11-20, 2016.

[20] M. Hijazi, A. Zeki, and A. Ismail, "Arabic Text Classification Using Hybrid Feature Selection Method Using Chi-Square Binary Artificial Bee Colony Algorithm," Computer Science, vol. 16, no. 1, pp. 213-228, 2021.

[21] A. J. Mohammed, Y. Yusof, and H. Husni, "Document clustering based on firefly algorithm," Journal of Computer Science, vol. 11, no. 3, pp. 453-465, 2015.

[22] A. J. Mohammed, Y. Yusof, and H. Husni, "GF-CLUST: A nature-inspired algorithm for automatic text clustering," Journal of Information and Communication Technology, vol. 15, no. 1, pp. 57-81, 2016.

[23] E. Cuevas, M. Cienfuegos, D. Zaldívar, and M. Pérez-Cisneros, "A swarm optimization algorithm inspired in the behavior of the social-spider," Expert Systems with Applications, vol. 40, no. 16, pp. 6374-6384, 2013.

[24] T. R. Chandran, A. Reddy, and B. Janet, "Text clustering quality improvement using a hybrid social spider optimization," International Journal of Applied Engineering Research, vol. 12, no. 6, pp. 995-1008, 2017.

[25] T. R. Chandran, A. Reddy, and B. Janet, "An effective implementation of social spider optimization for text document clustering using single cluster approach," in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018: IEEE, pp. 508-511.

[26] H. Rashaideh et al., "A grey wolf optimizer for text document clustering," Journal of Intelligent Systems, vol. 29, no. 1, pp. 814-830, 2020.

[27] L. M. Abualigah, A. T. Khader, M. A. Al-Betar, and M. A. Awadallah, "A krill herd algorithm for efficient text documents clustering," in 2016 IEEE symposium on computer applications & industrial electronics (ISCAIE), 2016: IEEE, pp. 67-72.

[28] L. M. Abualigah, A. T. Khader, and E. S. Hanandeh, "Hybrid clustering analysis using improved krill herd algorithm," Applied Intelligence, vol. 48, no. 11, pp. 4047-4071, 2018.

[29] P. S. More, B. S. Saini, and K. S. Bhatia, "Krill Herd (KH) algorithm for text document clustering using TF–IDF features," in Smart Computing: CRC Press, 2021, pp. 502-512.

[30] L. M. Abualigah and A. T. Khader, "Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering," The Journal of Supercomputing, vol. 73, no. 11, pp. 4773-4795, 2017/11/01 2017, doi: 10.1007/s11227-017-2046-2.

[31] M. Asif, A. A. Nagra, M. B. Ahmad, and K. Masood, "Feature Selection Empowered by Self-Inertia Weight Adaptive Particle Swarm Optimization for Text Classification," Applied Artificial Intelligence, pp. 1-16, 2021.

[32] A. Singh and A. Kumar, "Text document classification using a hybrid approach of ACOGA for feature selection," International Journal of Advanced Intelligence Paradigms, vol. 20, no. 1-2, pp. 158-170, 2021.

[33] R. Janani and S. Vijayarani, "Automatic text classification using machine learning and optimization algorithms," Soft Computing, vol. 25, no. 2, pp. 1129-1145, 2021.

[34] B. Diallo, J. Hu, T. Li, G. A. Khan, and A. S. Hussein, "Multi-view document clustering based on geometrical similarity measurement," International Journal of Machine Learning and Cybernetics, pp. 1-13, 2021.

[35] Y.-S. Lin, J.-Y. Jiang, and S.-J. Lee, "A similarity measure for text classification and clustering," IEEE transactions on knowledge and data engineering, vol. 26, no. 7, pp. 1575-1590, 2013.