International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 6, Issue 12, ISSN No. 2455-2143, Pages 52-58
Published Online April 2022 in IJEAST (http://www.ijeast.com)

A MULTI-OBJECTIVE GWO BASED CLUSTERING APPROACH FOR EFFICIENT DOCUMENT MINING

Darshna Rai
Assistant Professor, Department of Computer Science & Engineering
School of Research & Technology, People's University, Bhopal, MP, India
Email - darshna.rai@gmail.co

Dr Shailja Sharma
Associate Professor, Department of Computer Science & Engineering
Rabindtranath Tagore University Bhopal, Mendua 464993 MP, India
Email - shailja1901@rediffmail.com

Abstract: **Learning and data mining systems often use clustering. Such applications require efficient and fast clustering algorithms. However, clustering a dataset is often a difficult task, especially for complex datasets such as text classification. Clustering is a strategy of grouping data points based on their similarities. Data clustering can be done using any clustering algorithm. However, conventional partitioning clustering algorithms greatly depend on initial points and can drift towards inefficient clustering on improper initialization. Furthermore, these algorithms are less flexible for achieving multiple objectives required for clustering the complex overlapping classes. Therefore, in this work, we proposed a Gray Wolf Optimization (GWO) based multiobjective optimization algorithm for document clustering. GWO is a metaheuristic algorithm that works in the same way that grey wolves do when hunting prey. The proposed algorithm has the advantage of easily modifying clusters based on the objectives. The performance of the proposed algorithm on the Reuters-21578 dataset demonstrates that it outperforms k-means and affinity propagation algorithms.**

Keywords: **Text Mining, Clustering, Gray Wolf Optimization (GWO).**

I. INTRODUCTION

The rise of user-generated data on social media, microblogs, and e-commerce websites has resulted in a tremendous amount of data that may be used for different purposes [1].. However, different applications require different types of information, so it is necessary to cluster them according to the application requirements. This is done through document categorization. The most common method of document categorization is clustering.

Text clustering is a method for extracting classes, concepts, or groups of patterns from unstructured material automatically[2]. It attempts to organize or cluster an unstructured collection of things. As a result, the objects must be comparable to those in the same cluster while being distinct from those in different clusters. Biology, medicine, anthropology, marketing, and economics are just a few of the fields where clustering has been used [3].

As digitization techniques improve, a considerable portion of all written data is now kept digitally (as soft copies). Document clustering is thus one of the most essential uses, and it is becoming increasingly relevant [4]. A fast and efficient clustering method was required to get valuable data from a big database. Data mining and classical text mining share many similarities. Data mining employs a number of strategies to find the hidden information contained in underlying data structures. One of them is the clustering approach [5]. In the case of textual data, clustering algorithms seek methods to discover the underlying groupings so that a group (a set of documents) can form clusters with high similarities among individuals within the cluster and low similarities among individuals between clusters [6]. Traditional clustering approaches, such as k-means clustering, rely heavily on the initial choice of cluster center, which must be rerun many times to yield the best results.

These problems can be solved by treating the clustering as an optimization problem, much like, the clustering problem is generally defined as: Given a set of n patterns $X = \{x_1, x_2, \dots x_n\}$ in d dimensional space, partition the set X into k clusters $C = \{c_1, c_2, \dots c_k\}$that minimize a predetermined criterion (for example, sum of squared errors (SSE), entropy, f-measure, or accuracy) [4]. The advantage of meta heuristic clustering over classical clustering is that the former is unaffected by starting cluster locations and may be easily adjusted by user-defined objective functions.

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 6, Issue 12, ISSN No. 2455-2143, Pages 52-58
Published Online April 2022 in IJEAST (http://www.ijeast.com)

In this work, we proposed a new clustering approach based on Grey Wolf Optimization (GWO) [7], using intra-cluster and inter-cluster entropies as objectives. We combine these objectives into a single objective that minimizes when inter-cluster entropy increases and intra-cluster entropy decreases.
The rest of the paper is organized as in Section-II, a review of related work is presented. Section-III describes the GWO and objective functions. Section-IV presents the proposed algorithm, followed by the experimental results and analysis in Section-V. Finally, the conclusion is presented in Section-VI.

II. RELATED WORK

The most common meta-heuristic techniques used to tackle text document clustering problems that have been reported in the literature are discussed in this section.
Particle Swarm Optimization (PSO) is one of the powerful meta-heuristic optimization techniques utilized to handle clustering problems in the papers [8]. He-Nian et al. proposed OK-PSO for text clustering based on k-means (KM) and a PSO algorithm [9]. The KM is used to compute the distance between each word and the cluster centers. The 2-D Otsu algorithm was used to test the optimization of the clustering distance. To speed up threshold estimation, the technique employs the PSO algorithm [10]to search for the best threshold. The efficacy of the proposed strategy was tested using datasets and compared to other clustering approaches. The experimental findings demonstrated the superiority of the proposed method over the algorithms.
Song et al. [11] created a genetic algorithm based on ontology science, where the algorithm can solve text clustering problems in a self-organizing manner. They attempted to solve the text clustering problem by improving the evolutionary algorithm with a model known as latent semantic, which differs from the conventional vector space model in which each part of the text or vocabulary represents one dimension.
In various academic domains, ant colony optimization (ACO) has been applied to solve clustering difficulties [12]. ACO is used for multi-label text categorization with the relevance clustering classification algorithm in article[13].
Yang [14] proposes another fascinating meta heuristic algorithm the firefly algorithm (FA), which is inspired by the qualities of fireflies being attracted by the brightness of others. It was initially employed to solve the function optimization problem, and its performance is fairly impressive [15]. Levy flights were integrated with FA in [16] to improve its ability to disturb the solutions locally. In [17]an alternate method that leverages revised starting answers to increase the quality of the outcomes is presented.

III. MATHEMATICAL BACKGROUND

A. Gray Wolf Optimization

The Grey Wolf Optimization (GWO) was proposed by Mirjalili et al [7]. The GWO is influenced by grey wolf hunting behavior and social structure. The experimental findings demonstrated its capabilities and great performance in handling numerous classic engineering design challenges, such as spring tension, welded beam, and so on.

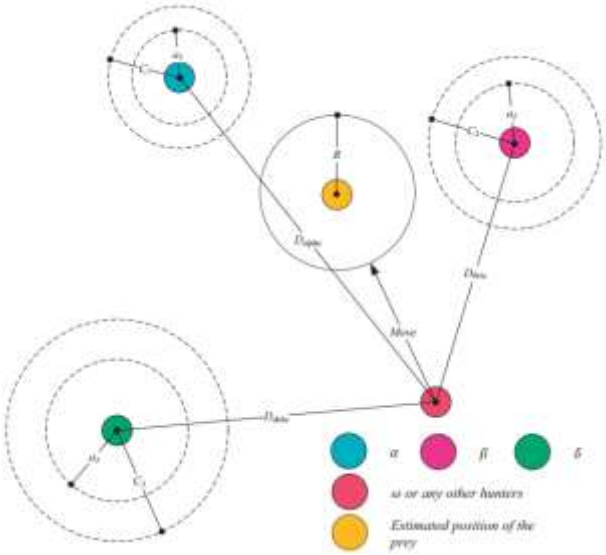

Figure 1: Movement of wolfs in GWO.

The GWO technique acknowledges the difficulty of finding the optimal answer as grey wolf prey hunting. The goal is the same as the best answer. Because the grey wolves' hunting method consists of three stages: encircling the prey, hunting, and attacking the prey, the GWO employs these strategies to discover the best answer. The grey wolves are specifically following the social order of leadership. The pack is led by the alpha (α) wolf, who remains at the top of the hierarchy.
Similarly, after alpha, beta (β) wolves are regarded the second level of wolves, while the third and fourth levels of wolves are known as delta (δ) and omega (ω), respectively. All wolves (alpha, delta, and omega) trail the alpha wolf, whereas delta and omega trail the beta wolves, and delta wolves trail omega. There are no followers because the omega remains at its lowest level.
The alpha, beta, and delta wolves lead the hunt, and the other wolves (omega) simply follow. The movement of the entire population in the optimization problem is driven by the top three best solutions, and these solutions are referred to as alpha, beta, and delta, respectively. The other solutions are referred to as omega. These places evolve until a solution is discovered; the entire process can be summarized as follows:

1. **Pray Encircling:** Gray wolves' first hunting maneuver is to encircle their prey. The encircling method of gray wolves during hunting is interpreted here as the population encircling the optimal solution. It is written in mathematical notation as :

$$\vec{D} = \left|\vec{C} \cdot \vec{X}_{prey}(j) - \vec{X}_i(j)\right|, \qquad (1)$$

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 6, Issue 12, ISSN No. 2455-2143, Pages 52-58
Published Online April 2022 in IJEAST (http://www.ijeast.com)

$$\vec{X}_i(j+1) = \vec{X}_{prey}(t) - \vec{A} \cdot \vec{D}, \qquad (2)$$

where j signifies the iteration number, $\vec{X}_{Prey}$ is pray position vector and $\vec{X}_i$ is i^{th} wolf position vector. The $\vec{A}$ and $\vec{C}$ are coefficient vectors and calculated as follows:

$$A = 2\vec{a} \cdot \vec{r}_1 - \vec{a}, \qquad (3)$$

$$\vec{C} = 2\vec{r}_2, \qquad (4)$$

where the magnitude of vector $\vec{a}$ is decreased linearly from 2 to 0 at each step and $\vec{r}_1, \vec{r}_2$ are random vectors in the range of 0, to 1

2. **Hunting:** The prey position is known in the real hunting situation, but the optimal solution is unknown in the optimization issue, so alpha, beta, and delta positions are used to derive a rough location of the optimum solution. The wolves' location has been revised as follows:

$$\vec{X}_{wolf}(i+1) = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3}{3}, \qquad (5)$$

The values of $\vec{X}_1, \vec{X}_2$ *and* $\vec{X}_3$ are the assumed coarse location of the optimum solution ($\vec{X}_{prey}$) based on α, β and δ wolfs and are calculated as:

$$\vec{X}_1 = \vec{X}_\alpha - \vec{A}_1 \cdot (\vec{D}_\alpha), \qquad D_\alpha = |\vec{C}_1 \cdot \vec{X}_\alpha - \vec{X}_i|, \qquad (6)$$

$$\vec{X}_2 = \vec{X}_\beta - \vec{A}_2 \cdot (\vec{D}_\beta), \qquad D_\beta = |\vec{C}_1 \cdot \vec{X}_\beta - \vec{X}_i|, \qquad (7)$$

$$\vec{X}_3 = \vec{X}_\delta - \vec{A}_3 \cdot (\vec{D}_\delta), \qquad D_\delta = |\vec{C}_1 \cdot \vec{X}_\delta - \vec{X}_i|, \qquad (8)$$

3. **Attacking:** As the gray wolf tightens its hold on the victim, the prey's movement grows less and smaller as the wolf advances, until the prey eventually stops moving and the wolf executes the final attack. This scenario is reproduced in the mathematical model by decreasing the values of a vector $\vec{a}$. Linearly from 2 to 0 with each iteration, limiting the search region for optimum location (solution) and population movement sites and gradually arriving at the optimum position.

B. GWO Clustering

To increase the performance of the text clustering technique, we offer a novel formula that incorporates two separate metrics as objective functions to make an accurate judgment throughout the clustering process. As previously stated, this combination comprises of two independent measures that are commonly utilized in the text clustering domain separately, namely, intra-cluster entropy as Eq. (10) and inter-cluster entropy as Eq. (11). We combine both the equations Eq. (10) and Eq. (11) into Eq. (9) in order to maximize the benefits of both.

$$F_{obj} = w_A H(X) + w_B H^*(Y) \qquad (9)$$

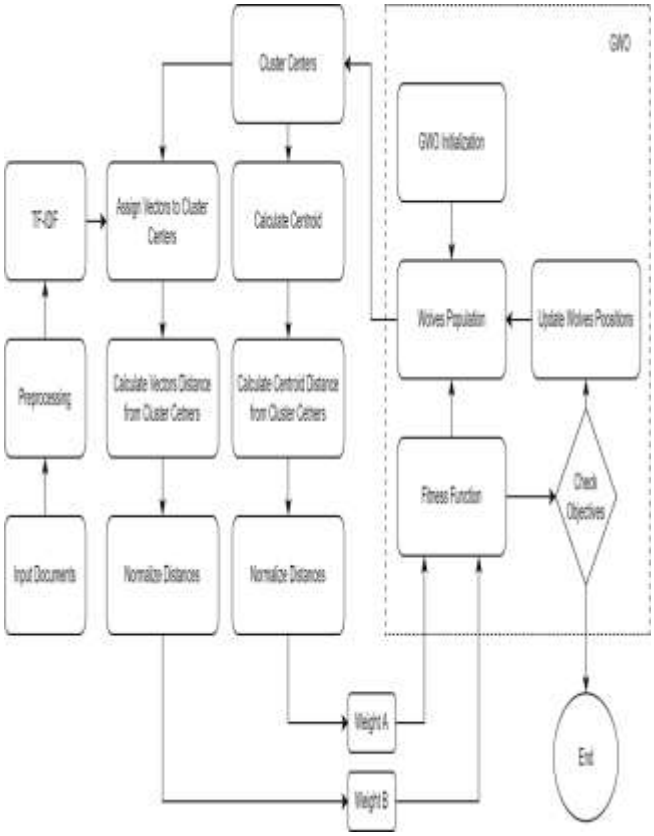

Figure 1: Illustration of the Proposed Approach.

A. Entropy

Because entropy defines the distribution of data, it can be used to assess the quality of clusters. The first step in calculating the entropy is to compute the $p_i(x_j)$, which signifies the probability of existing a member at position x_j of the cluster i. The entropy for cluster i(intra-cluster) is then determined using the conventional formula:

$$H_i(X) = -\sum_{x_j \in X} p_i(x_j) \cdot log\big(p_i(x_j)\big) \qquad (10)$$

Where the total members is denoted by X. Then total entropy is derived by executing the following weighted sum of individual entropies:

$$H(X) = \sum_{i=1}^{N} \frac{n_i \cdot E_i(X)}{n} \qquad (11)$$

Where n and n_i denotes the total data points, and the data points in i^{th} cluster.

Similarly the inter-cluster entropy is calculated by:

$$H_i^*(Y) = -\sum_{y \in Y^*} p(y_j^*) \cdot log\big(p(y_j^*)\big) \qquad (12)$$

Where $p(y_j^*)$ denotes the probability of the j^{th} cluster center from the mean of $Y^* = \{y_1^*, y_2^*, \dots, y_N^*\}$, where, N is the total number of clusters.

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 6, Issue 12, ISSN No. 2455-2143, Pages 52-58
Published Online April 2022 in IJEAST (http://www.ijeast.com)

IV. REUTERS-22173 DATASET DESCRIPTION

Several published studies have used the Reuters-22173 test collection since it was made available, and we believe that the Reuters-21578 collection will be even more valuable. There are 22 files in the Reuters-21578 collection. In the first 21 files (reut2-000.sgm through reut2-020.sgm), there are 1000 documents, while the last file (reut2-021.sgm) contains 578 documents[18].

This package includes an SGML DTD for specifying the data file format and six files for specifying the categories for indexing the data. In the Reuters-21578 collection, the documents are newswire stories from Reuter, and the categories are five content-related sets.

A human indexer determined which categories and sets each document belonged to[19], the indexed categories are listed in Table-I.

V. PROPOSED ALGORITHM

In Fig. 2 an illustrative block diagram of the proposed algorithm is shown, which can be described as follows:

Step 1 Reading Dataset: The first step of the algorithm is to read the Routers dataset and create a unique identifier for the each document. For the i^{th} document the identifier is represented by the D_i while the whole set of identifier is denoted by $\boldsymbol{D}$.

Step 2 Grouping by Category: As the dataset also contains the label for each document according to their category as presented in Table 1. The documents identifiers are also grouped according to these labels which are later used for the cross validation to evaluation of the performance of the proposed algorithm.

Step 3 Preprocessing: In this step the documents are prepared for the feature extraction. To achieve this following operations are performed:

1 Tokenization
2 Stop words removal
3 Stemming
4 Lemmatizing

Step 4 Documents Representation: document representation projects (or represents) each document into feature space, by representing the document through a feature space vector (or feature vector). In the presented work TF-IDF [20]is utilized. The feature vector for the documentsD_i is denoted by $V_i \in \mathbb{R}^{m \times n}$ (here m and n denoted the length of feature vector and total number of documents in dataset respectively) and the set of feature vectors belongs to documents in D^G is denoted by $V_i^G \in \mathbb{R}^{m \times n_i}$ (where n_i total documents in group G_i).

Step 5 GWO Clustering: the documents representing feature vectors $\mathcal{V}$ are grouped using the GWO clustering algorithm as described in subsection A and B of Section III. The algorithm is forced for six clusters as the dataset originally had the same number of clusters. After the clustering, the documents belonging to i^{th} cluster is represented by $\mathcal{D}^i$.

Step 6 Performance Evaluation Metrics: to validate and compare the performance of the algorithm the quality of the clusters is evaluated using these metrics as described in Section-VI.

VI. PERFORMANCE EVALUATION METRICS

In this work following measures are used to evaluate the proposed clustering algorithm.

B. Entropy

Since entropy describes the distribution of data, it can be used as the quality measure of the clusters. To estimate the entropy the first step is to calculate the p_{ij} which denotes the probability of an element of cluster j belongs to class i. Then the entropy for cluster j is calculated using the following standard formula:

$$E_j = -\sum_{i \in N} p_{ij} \cdot log(p_{ij}) \tag{13}$$

Where N denotes the total number of classes. Similarly, total entropy is calculated by performing a weighted sum of individual entropies as follows:

Table 1: Description of Reuters-21578 dataset groups.

Category Group	Number of Elements	1+ Occurrences	20+ Occurrences
Exchanges	39	32	7
Orgs	56	32	9
People	267	114	15
Places	175	147	60
Topics	135	120	57

$$E_C = \sum_{j=1}^{N} \frac{n_j \cdot E_j}{n} \tag{14}$$

Where n and n_j denotes the total data points, and the data points in j^{th} cluster.

C. Precision, Recall and, F-Measure

Precision defines the classes distribution in a cluster andis given as:

$$\text{Precision}(i, j) = \frac{n_{ij}}{n_j} \tag{15}$$

Where n_j and n_{ij} denotes total data point in j^{th} clusterand data points in j^{th} cluster belongs to i^{th} class.

Recall defines a class distribution over clusters and is given as:

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 6, Issue 12, ISSN No. 2455-2143, Pages 52-58
Published Online April 2022 in IJEAST (http://www.ijeast.com)

$$\text{Recall}(i,j) = \frac{n_{ij}}{n_i} \quad (16)$$

Where n_i denotes total data points belonging to i^{th} class.
F-measure is the harmonic mean of Precision and Recall and is given as:

$$F(i,j) = 2 \times \frac{\text{Precision}(i,j) * \text{Recall}(i,j)}{\text{Precision}(i,j) + \text{Recall}(i,j)} \quad (17)$$

For total clusters, the F-Score is calculated as:

$$F = \sum_{i=1}^{N} \frac{n_i}{n} \max\{F(i,j)\} \quad (18)$$

Where n is the total number of data points.

D. Purity

Evaluate whether each cluster contains only examples from the same class:

$$PU = \sum_{i=1}^{N} p_i \left(\max_j \frac{p_{ij}}{p_i} \right) \quad (19)$$

Where $p_i = \frac{n_i}{n}$, $p_{ij} = \frac{n_{ij}}{n}$ and other terms are the same as defined in Eq. (15) and (16).

VII. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed algorithm a rigorous analysis was performed. The analysis involves the evaluation of performance using different quality measures and different dataset lengths. Finally, the algorithm is also compared with some state of art algorithms. These results are presented from Table-2 to Table-7.

Table 2: Performance Evaluation Results for Precision.

Dataset Length Percentage of Total	Precision			
	K-Means	Fuzzy C-Means	AP	Proposed
25	0.53	0.56	0.64	0.70
50	0.50	0.55	0.59	0.70
75	0.48	0.48	0.53	0.65
100	0.39	0.43	0.52	0.60

Table 3: Performance Evaluation Results for Recall.

Dataset Length Percentage of Total	Recall			
	K-Means	Fuzzy C-Means	AP	Proposed
25	0.56	0.58	0.59	0.72
50	0.49	0.51	0.52	0.65
75	0.42	0.40	0.41	0.62
100	0.32	0.33	0.36	0.60

Table 4: Performance Evaluation Results for F-Score.

Dataset Length Percentage of Total	F-Measure			
	K-Means	Fuzzy C-Means	AP	Proposed
25	0.55	0.57	0.61	0.71
50	0.50	0.53	0.55	0.67
75	0.45	0.44	0.46	0.63
100	0.36	0.37	0.43	0.60

Table 5: Performance Evaluation Results for Entropy.

Dataset Length Percentage of Total	Entropy			
	K-Means	Fuzzy C-Means	AP	Proposed
25	0.43	0.36	0.25	0.20
50	0.45	0.43	0.24	0.22
75	0.55	0.36	0.28	0.23
100	0.58	0.38	0.30	0.24

Table 6: Performance Evaluation Results for Purity.

Dataset Length Percentage of Total	Purity			
	K-Means	Fuzzy C-Means	AP	Proposed
25	0.52	0.63	0.69	0.76
50	0.51	0.63	0.68	0.75
75	0.55	0.64	0.66	0.73
100	0.56	0.57	0.62	0.71

Table 7: Performance Evaluation Results for Time.

Dataset Length Percentage of Total	Time (Seconds)			
	K-Means	Fuzzy C-Means	AP	Proposed
25	32	46	85	90
50	45	63	254	133
75	62	94	440	297
100	92	130	640	436

Tables 2 through 4 compare Precision, Recall, and F-Score measurements for various algorithms.
These results show that as the size of the dataset grows larger, the performance of all algorithms declines. Furthermore, for each dataset size, the suggested technique outperforms competing algorithms.
Furthermore, for a dataset size of 25%, the suggested approach obtains the maximum Precision, Recall, and F-Score of 0.70, 0.72, and 0.71, respectively, which reduces to 0.62, 0.62, and 0.60 for a dataset size of 100%.
Further examination of the findings shown in Tables 2–4 reveals that the suggested algorithm achieves the greatest

improvements in Precision, Recall, and F-Score of 9.37 percent, 22.03 percent, and 16.39 percent, respectively.
Tables 5 and 6 provide the assessment results for Entropy and Purity, which show that the suggested algorithm improves Entropy by 33.33 percent maximum and Purity by 10.14 percent maximum.
Finally, Table-7 shows the time performance of all algorithms. According to the results, K-Means is the fastest performing algorithm, followed by Fuzzy C-Means (FCM). Although the suggested technique takes longer than K-Means and FCM, it is faster than Affinity Propagation (AP).

VIII. CONCLUSION

This study presents a text clustering implementation based on GWO. The suggested approach is thoroughly analyzed using different evaluation criteria and compared to some state-of-the-art algorithms.
Overall, the findings reveal that the proposed method outperforms the compared algorithms in terms of both quality and processing time. This boost in performance is accomplished by effectively arranging the documents according to the generated objective function.

IX. REFERENCES

[1] E. Souza, D. Santos, G. Oliveira, A. Silva, and A. L. I. Oliveira, "Swarm optimization clustering methods for opinion mining," Nat Comput, vol. 19, no. 3, pp. 547–575, Sep. 2020, doi: 10.1007/s11047-018-9681-2.

[2] L. Abualigah et al., "Advances in Meta-Heuristic Optimization Algorithms in Big Data Text Clustering," Electronics, vol. 10, no. 2, Art. no. 2, Jan. 2021, doi: 10.3390/electronics10020101.

[3] A. Ghosal, A. Nandy, A. K. Das, S. Goswami, and M. Panday, "A Short Review on Different Clustering Techniques and Their Applications," in Emerging Technology in Modelling and Graphics, Singapore, 2020, pp. 69–83. doi: 10.1007/978-981-13-7403-6_9.

[4] C.-W. Tsai, W.-C. Huang, and M.-C. Chiang, "Recent Development of Metaheuristics for Clustering," in Mobile, Ubiquitous, and Intelligent Computing, Berlin, Heidelberg, 2014, pp. 629–636. doi: 10.1007/978-3-642-40675-1_93.

[5] A. Saxena et al., "A review of clustering techniques and developments," Neurocomputing, vol. 267, pp. 664–681, Dec. 2017, doi: 10.1016/j.neucom.2017.06.053.

[6] C. D. Tupper, "13 - Concepts of Clustering, Indexing, and Structures," in Data Architecture, C. D. Tupper, Ed. Boston: Morgan Kaufmann, 2011, pp. 241–253. doi: 10.1016/B978-0-12-385126-0.00013-9.

[7] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey Wolf Optimizer," Advances in Engineering Software, vol. 69, pp. 46–61, Mar. 2014, doi: 10.1016/j.advengsoft.2013.12.007.

[8] L. M. Abualigah, A. T. Khader, and E. S. Hanandeh, "A new feature selection method to improve the document clustering using particle swarm optimization algorithm," Journal of Computational Science, vol. 25, pp. 456–466, Mar. 2018, doi: 10.1016/j.jocs.2017.07.018.

[9] H.-N. Chen, B. He, L. Yan, J. Li, and W. Ji, "A Text Clustering Method Based on Two-Dimensional OTSU and PSO Algorithm," in 2009 International Symposium on Computer Network and Multimedia Technology, Jan. 2009, pp. 1–4. doi: 10.1109/CNMT.2009.5374525.

[10] J. Kennedy and R. Eberhart, "Particle swarm optimization," in Proceedings of ICNN'95 - International Conference on Neural Networks, Nov. 1995, vol. 4, pp. 1942–1948 vol.4. doi: 10.1109/ICNN.1995.488968.

[11] W. Song, C. H. Li, and S. C. Park, "Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures," Expert Systems with Applications, vol. 36, no. 5, pp. 9095–9104, Jul. 2009, doi: 10.1016/j.eswa.2008.12.046.

[12] M. A. Shi-xia, L. I. U. Dan, and J. I. A. Shi-jie, "Text Clustering Algorithm Based on Ant Colony Algorithm," vol. 36, no. 8, pp. 206–207, Apr. 2010, doi: 10.3969/j.issn.1000-3428.2010.08.072.

[13] P. Nema and V. Sharma, "Multi-label text categorization based on feature optimization using ant colony optimization and relevance clustering technique," in 2015 International Conference on Computers, Communications, and Systems (ICCCS), Nov. 2015, pp. 1–5. doi: 10.1109/CCOMS.2015.7562842.

[14] X.-S. Yang, "Firefly Algorithms for Multimodal Optimization," in Stochastic Algorithms: Foundations and Applications, Berlin, Heidelberg, 2009, pp. 169–178. doi: 10.1007/978-3-642-04944-6_14.

[15] X.-S. Yang and S. Deb, "Eagle Strategy Using Lévy Walk and Firefly Algorithms for Stochastic Optimization," in Nature Inspired Cooperative Strategies for Optimization (NICSO 2010), J. R. González, D. A. Pelta, C. Cruz, G. Terrazas, and N. Krasnogor, Eds. Berlin, Heidelberg: Springer, 2010, pp. 101–111. doi: 10.1007/978-3-642-12538-6_9.

[16] X.-S. Yang, "Firefly Algorithm, Levy Flights and Global Optimization," arXiv:1003.1464 [math], Mar. 2010, Accessed: Mar. 30, 2022. [Online]. Available: http://arxiv.org/abs/1003.1464

[17] S. Łukasik and S. Żak, "Firefly Algorithm for Continuous Constrained Optimization Tasks," in Computational Collective Intelligence. Semantic Web, Social Networks and Multiagent Systems, Berlin, Heidelberg, 2009, pp. 97–106. doi: 10.1007/978-3-642-04441-0_8.

[18] R. Feldman and I. Dagan, "Knowledge Discovery in Textual Databases (KDT)," p. 6.

[19] J. M. G. Hidalgo and M. de B. Rodriguez, "Integrating a Lexical Database and a Training Collection for Text Categorization," arXiv:cmp-lg/9709004, Sep. 1997, Accessed: Dec. 06, 2021. [Online]. Available: http://arxiv.org/abs/cmp-lg/9709004

[20] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries."

[21] S. Tsuge, M. Shishibori, S. Kuroiwa, and K. Kita, "Dimensionality reduction using non-negative matrix factorization for information retrieval," in 2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat.No.01CH37236), Oct. 2001, vol. 2, pp. 960–965 vol.2. doi: 10.1109/ICSMC.2001.973042.