Volume 11, Issue 4
August 2022

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

An Automated Efficient Document Clustering Method using Hybrid Fitness Distance Balance Based Coyote Optimization and Capuchin Search-Based Neural Network Classifier

[1]Perumal Pitchandi

[1]Department of Computer Science and Engineering
Sri Ramakrishna Engineering College, Coimbatore
E-mail: perumalp@srec.ac.in

ABSTRACT

Clustering of high-dimensional document datasets is one of the paramount challenges in text mining. Traditional cluster algorithms have undergone many conceptual and algorithmic changes recently, as well as several concepts such as ambiguous theory, ballet intelligence, genetic algorithm and oncology to enhance performance. In this paper, propose an automated efficient document clustering method using a hybrid fitness-distance balance (FDB) based coyote optimization and capuchin search-based neural network classifier (AEDC). The proposed method process consists of three tires. First, introduce a modified ant lion optimization (MALO) algorithm for data pre-processing which removes the unwanted artifacts and redundant content from the documents. Collect 'n' number of features from the pre-processed documents but all features are not required for the clustering. Second, illustrate the FDB-based coyote optimization (FDB-CO) algorithm for optimal feature selection which computes the best and optimal features among multiple features in the document. Then, offer a capuchin search-based neural network (CSNN) for optimal document clustering which improves the performance of clustering. Finally, the proposed method evaluates using two datasets, namely the Reuter database, 20 Newsgroups database and own dataset. The performance of the proposed AEDC method can compare with existing methods in terms of accuracy, precession, recall, F-measure and Rand index.

Keywords: *Document Clustering, Neural Network Classifier, Ant Lion Optimization, Capuchin Search*

1. INTRODUCTION

Document clustering is the use of cluster analysis in text documents for automated document systems, title recovery, faster data recovery or filtering. Document taxonomy is one of the most important processes of web classification and text extraction, and clustering is an intuitive technology for automated document taxonomy. Using blurred members to measure the size of objects compared to clusters creates a natural way for blurred cluster files to shoot at each other between clusters. Various representative models have been developed for documents, including synthetic models, semantic models with external dynamics such as Word and Wikipedia, and models with additional information such as editor and journal scientific article, to effectively maintain adequate information of the original data. Document abstraction techniques have received a lot of attention as a guide to documents and an effective way of compiling. To find out the forms and key features, a cluster study of medical documentation was conducted to compile the medical documentation into one major cluster. The cluster group data overseen by the clustering algorithm contains 17 clinical references, showing that different medical domains use different semantic and semantic forms. The linguistic features of medical notes from different organizations were compared and medical specialty was found throughout the company using clustering technology of documents. Cluster cryptographic notation was used for medical references, and latent semantic indexing is an excellent way to measure the similarity of clinical references. Nine semantic similarities of ontology-based terms were evaluated to summarize medical documents. The evaluation results of combining the names of symptoms/drugs with cluster medical documentation.

The vector space model (VSM) is the primary way to present a document in a wide range of applications, where each file is a vector and each attribute is a unique term that appears in the document. A moderately sized document database is converted to thousands of dimension feature locations, where the distance between any two vectors is very similar. This makes high levels of clustering an issue that cannot be properly addressed by cluster approaches. Attempts have been made in the literature to improve clustering approaches, especially documentation based on document-based remote measurements or similar measures, and new cluster models in line with high-volume document data. Spherical clustering, co-clustering, and negative non-matrix factorization are multiple representatives. Although these specially designed clustering approaches work well in document classification, many believe that the target database can be fully loaded into memory, which significantly limits the practical use of large-scale document processing. Although the main feature of a single pass is the continuous processing of the improved cluster, the Divide-Ensemble method is used exclusively for objects from different clusters. All three

schemes are combined with dim clusters to achieve extended dimensions. Existing large-scale ambiguous cluster approaches are effective in solving large-scale problems in many programs. Data collection is done for many purposes and is not required for machine learning. Therefore, it is necessary to identify and obtain relevant data for a given analytical purpose. Each learning system has specific requirements on how to present data for analysis, so the data should be modified to meet these requirements. In addition, the selection of specific data to be analyzed will significantly increase the number of studied models. For these reasons, data preparation is an important part of any machine study. Data processing is the most time-consuming part of any car training program. Incorrect information was given in the documents; the use of ambiguous set theory is suggested. Blurred C-algorithms and obscure hierarchical clustering algorithms were used for the document cluster. To eliminate these shortcomings, ant-based obscure clustering algorithms and clustering algorithms that can handle an unknown number of clusters are known as obscure Ks. Based on the model of vector space, the comparison between the two documents is measured by the vector distance, i.e. the Manhattan distance and the Euclidean distance. These methods have no environmental significance. Ontology-based ambiguous Document Clusters were used for documents with limited subgroups of selected terms based on limited ontology. These methods control application domains that are difficult to generalize without proper domain ontology. For further improvement, an automated efficient document clustering method is proposed based on hybrid FDB-based coyote optimization and capuchin search-based neural network classifier.

The main contributions of the proposed method are summarized as follows:

1. A modified ant lion optimization (MALO) algorithm is used for data pre-processing which removes the unwanted artifacts and redundant content from the documents.
2. Collect ‘n’ number of features from the pre-processed documents but all features are not required for the clustering. FDB-based coyote optimization (FDB-CO) algorithm for optimal feature selection, it computes the best and optimal features among multiple features in the documents.
3. Capuchin search-based neural network (CSNN) is used for optimal document clustering which improves the performance of clustering.

The rest of the paper is organized as follows: Sect. 2 describes the recent works related to efficient document clustering methods. Sect. 3 provides the problem methodology and system model of the proposed AEDC technique. Sect. 4 gives the working model of the proposed AEDC technique with the corresponding mathematical analysis. Then, simulation results of proposed and existing techniques are discussed in Sect. 5. Finally, the paper concludes in Sect. 6.

2. RELATED WORK

SHERI et al. [21] have proposed an automatic synchronization build function based on a text classifier. Creating the main partitions of documents the main partitions are created by two different DIMs. A similar building measurement system creates a database for identifying clusters of identifiable documents and creating a text classifier. Yoon et al. [22] have proposed a probabilistic probability network map, probability output model and calculation method for superior document clustering. Additionally, network-based neurons have developed a new way of reflecting the strength of document relationships based on document ranking, given the importance of downloading files related to external files. Given the importance of the document, clusters of continuous documents can be focused on and presented as representative documents for chores such as data presentation and data analysis. Yarlagadda et al. [23] have proposed that input documents are pre-processed and extracted based on TFIDF and WordNet features. After the completion of extraction, knowledge of characteristic features is often built into the materials. Finally, the documents are compiled using a specific Rn-MSA, which is a combination of the Rider Optimization Algorithm (ROA) and the Moth Search Algorithm (MSA). Document cluster evaluation based on specific Motsup and RN-MSA is evaluated based on accuracy, assignment, f-measurement, and accuracy. Bikku et al. [24] have proposed that big data applications such as bioinformatics and DNA sequencing require the development of efficient taxonomies with high performance. Two major issues that need to be legally addressed for efficient and robust extraction of biomedical data are dimensional component space management and high-efficiency accuracy. Curiskis et al. [25] have proposed to evaluate different document clustering and title modelling techniques for the three sources of Twitter and Reddit data. It combined the representation of four different characteristics obtained from the Time Frequency Inverting Document Frequency (TF-IDF) Metric and Speech Tab models by the 4 clustering methods and included a hidden trick distribution title model for comparison. Different evaluation procedures are used in the literature, so it provides a discussion and recommendation of the most appropriate external measures for this task. It also shows the effectiveness of methods for data sets of different lengths.

Abasi et al. [26] have proposed a new method for modifying the MVO algorithm called the link-based Multi-verse optimizer algorithm (LBMVO) to improve the performance of the original MVO. The enhancement is that the MVO algorithm incorporates a neighbourhood operator to enhance the search capability through the new probability factor neighbourhood selection strategy (NSS). In addition to the five static databases used in the data cluster domain, the performance of the proposed LBMVO was tested on six static databases used in the text cluster domain. Park et al. [27] have proposed a conceptually simple but experimental cluster framework called advanced document clustering (ADC). It is designed to increase synthetic and semantic importance by gaining structural features and using cluster modules. In the ADC cluster module, semantic similarity can be measured using cosine similarity and centroid updating, while centroids are determined at the mini-module input. In addition, this

makes less use of cross-entropy losses because the training program is biased if the model settings are incorrect. AlMahmoud et al. [28] have proposed Clustering Arabic Documents based on Bond Energy (CADBE), which attempts to install and display natural variable clusters on large volumes of data. CADBE has three stages in compiling Arabic documents: the initial step is to establish a cluster association matrix with BE, the second step is to automatically divide the cluster matrix into smaller synchronous clusters using a new and innovative method, and the final step is to obtain a cluster.

Huang et al. [29] have proposed a multi-source document clustering model, namely, the Hierarchical Dirichlet Multinomial Allocation (HDMA) model, to solve all the above problems. The HDMA model explores two stages of creating a topic, and title sources learn to share their common characteristic features with the data source while retaining the local characteristics of the title source. Each data source is used in a separate title to determine the value of the source level title. Kim et al. [30] have proposed improvements on spherical k-means to overcome these issues during document clustering. Not only does our specific boot system guarantee scattered start points, but it's also 1000 times faster than the previously known boot modes k-means clustering. Additionally, activate the frequency using centrifugal vectors using database drives that control its value by cluster. Additionally, a supervised cluster labelling system that effectively provides important keywords to describe each cluster. The research gap summary is given in Table 1.

Table 1 **Summary of research gap**

Ref	**Algorithm**	**Methodology**	**Parameters**
21	k-means algorithm	DIM method	F-Measure, precision
22	k-means algorithm	Feature selection	Accuracy, recall
23	Rn-MSA algorithm	pruning techniques	precision, recall,
24	Machine learning algorithm	Feature selection	Sensitivity
25	clustering algorithms	online social networks	accuracy
26	MVO algorithm	TDC methods	Recall
27	*k* -means algorithm	ADC methods	F-Measure, accuracy
28	Bond Energy Algorithm	machine learning methods	Accuracy
29	Gibbs sampling algorithm	HDMA model,	precision
30	k-means algorithm	cluster labelling method	F-Measure

3. PROBLEM METHODOLOGY AND SYSTEM ARCHITECTURE

MapReduce is proposed to change the K-Means cluster algorithm based on the graph reductions for the document data set. The uniqueness of the study lies in the fact that the algorithm is studied in detail using several experiments. Although each experiment uses the same specific algorithm, the tests are designed based on cluster size, data flow volume, and associated operation time. The performance time obtained in a particular K-fish is comparable to the regular performance of a K-fish. The results show that this algorithm is more efficient than traditional Q-tools for clustering datasets of all sizes. Experiments have shown that graphical redundancy clustering works most efficiently as data set size and Hadoop cluster size increase. Distributed Document Clustering Algorithms Clustering is based on the availability of distributed resources [31].

High-dimensional document data collection is a major challenge in extracting clustering texts. Traditional cluster algorithms have undergone many conceptual and algorithmic changes in recent years, including ambiguous theory, coronary intelligence, genetic algorithms, oncology, word networking, word separation, and multiple algorithmic performance achievements. The algorithm, along with the latest developments in conceptual changes, also requires an extended clustering framework to process distributed large document data sets. This is due to the proliferation of large numbers of documents and the inability of central supercomputers to process large volumes of documents.

To solve above-mentioned problem, propose an automated efficient document clustering (AEDC) method using hybrid FDB-based coyote optimization and capuchin search-based neural network classifier. The working function of the proposed AEDC method is shown in Fig. 1.

Input Documents

Volume 11, Issue 4
August 2022

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

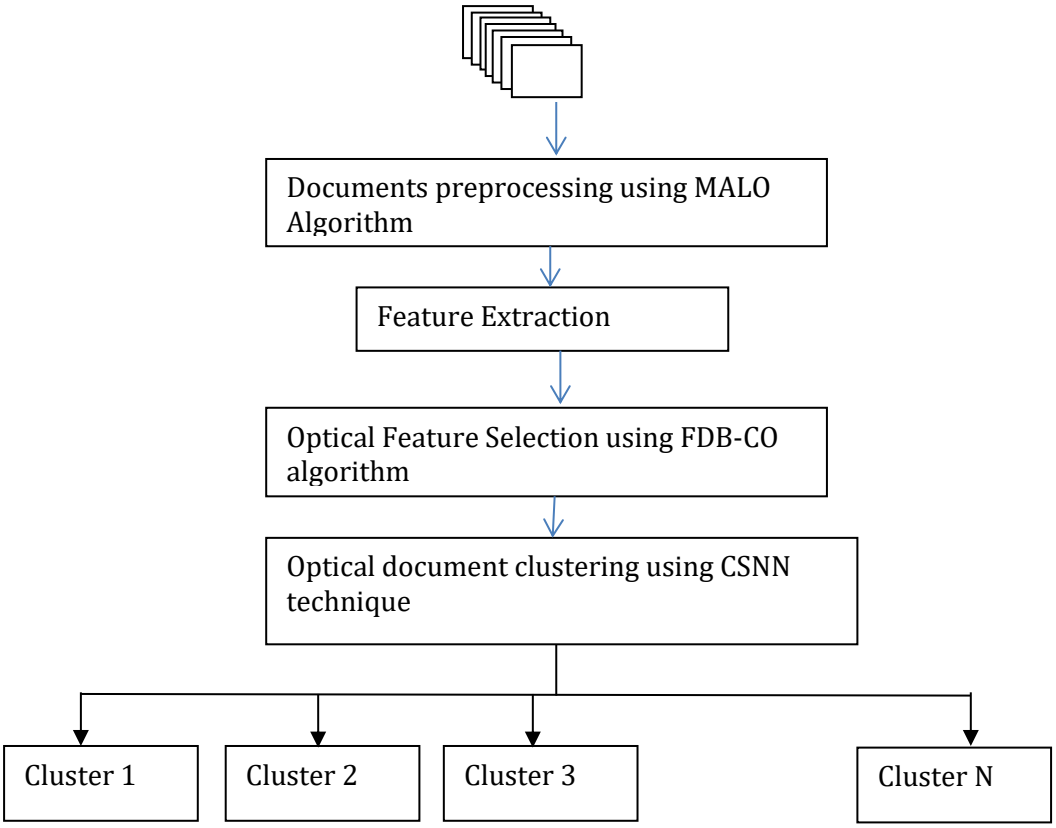


Fig. 1 System architecture of proposed AEDC method

4. PROPOSED AUTOMATED EFFICIENT DOCUMENT CLUSTERING (AEDC) METHOD

4.1 Pre-processing using modified ant lion optimization algorithm (MALO)

Document summarization is essential to gain knowledge from these unnamed lessons, which take place after several stages of pre-processing. Pre-processing often removes words. Another way to get information is to group processing documents into text. Pre-processes have a big impact on the success of knowledge acquisition. The database contains a variety of data collected from a variety of data sources. Because of this diversity, the actual data is inconsistent and serious. If the data do not match, the mining process can lead to confusion, which can lead to erroneous results. Pre-processing of this data is used to obtain consistent and accurate data. Here document clustering is done using a modified ant lion optimization algorithm (MALO). At the larval stage, they usually eat ants. Then he disappears to the bottom of the corner and waits. When the ant starts to get trapped, it starts trapping in the prey net. After catching a predator, Random gates are used to model the ant movement random search engine:

$$Y(T) = [0, Cumsum(2s(T_1) - 1), Cumsum(2s(T_m) - 1), \ldots, Cumsum(2s(T_m) - 1)] \quad (1)$$

Where Y (T) is the random gate of the ants, m is the maximum amount of iterations, T denotes the present frequency, Cumsum denotes whole sum, and s (T) is the constant function, which is described as follows:

$$s(T) = \begin{cases} 1 & Rand > 0.5 \\ 0 & otherwise \end{cases} \quad (2)$$

Rand denotes a random number in the range [0, 1]. The size of ants and ant lions then gives the matrix Eq. 3 and Eq. 4 respectively.

Volume 11, Issue 4
August 2022

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

$$N_{ant} = \begin{bmatrix} B_{1,1} & B_{1,2} & \dots & \dots & B_{1,c} \\ B_{2,1} & B_{2,2} & \dots & \dots & B_{2,c} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ B_{m,1} & B_{m,2} & \dots & \dots & B_{m,c} \end{bmatrix} \quad (3)$$

$$N_{antlion} = \begin{bmatrix} BW_{1,1} & BW_{1,2} & \dots & \dots & BW_{1,c} \\ BW_{2,1} & BW_{2,2} & \dots & \dots & BW_{2,c} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ BW_{m,1} & BW_{m,2} & \dots & \dots & BW_{m,c} \end{bmatrix} \quad (4)$$

where N_{ant} is a team that defends the status of ants, $N_{antlion}$ is A team that maintains antlion status, $B_{j,i}$ gives the value of the ant j by the i-th number, the total number of ants is denoted by m, and the number of dimension is denoted by c. An exercise function is used to evaluate each ant and ant lion. The results are stored in a matrix as Equation 5 and 6.

$$N_{oB} = \begin{bmatrix} g([B_{1,1,}\ B_{1,2},\dots,B_{1,c}]) \\ g([B_{2,1,}\ B_{2,2},\dots,B_{2,c}]) \\ \vdots \\ \vdots \\ \vdots \\ g([B_{m,1,}\ B_{m,2},\dots,B_{m,c}]) \end{bmatrix} \quad (5)$$

$$N_{oBW} = \begin{bmatrix} g([BW_{1,1,}\ BW_{1,2},\dots,BW_{1,c}]) \\ g([BW_{2,1,}\ BW_{2,2},\dots,BW_{2,c}]) \\ \vdots \\ \vdots \\ \vdots \\ g([BW_{m,1,}\ BW_{m,2},\dots,BW_{m,c}]) \end{bmatrix} \quad (6)$$

where N_{oB} is a ants exercise team, N_{oBW} is a team that for ant lion fitness, $B_{j,i}$ gives The value of the j ant on the i scale, the amount of ants is m, and g is denoted as objective function. The ant accidentally moves to the search area, Eq 8 is used to normalize the condition of ants.

$$Y_j^T = \frac{(Y_j^T - b_j)(c_j^T - d_j^T)}{c_j - b_j} + d_j^T \quad (7)$$

Where b_j, a_i is The minimum random gates and maximum random gates of variable j respectively, d_j^T, c_j^T is With variable frequency of minimum and maximum iteration T. The following equations are presented to model the behavior of the viewer trap

$$d_j^T = antlion_j^T + d^T \quad (8)$$

$$c_j^T = antlion_j^T + c^T \quad (9)$$

$$d^T = \frac{d^T}{J} \quad (10)$$

$$c^T = \frac{c^T}{J} \quad (11)$$

where d^T, c^T the variables are minimum and maximum when T is repeated respectively, d_j^T, c_j^T is the variables of the first piece of ant are ith, respectively, minimum and maximum, $antlion_i^T$ When T repeats, the i ant indicates the position of the lion, and J is The sliding ratio is as follows:

$$J = \begin{cases} 1+10^6\ Iter/\max iter & if \quad 0.95\max iter < Iter < \max iter \\ 1+10^5\ Iter/\max iter & if \quad 0.90\max iter < Iter < 0.95\max iter \\ 1+10^4\ Iter/\max iter & if \quad 0.75\max iter < Iter < 0.90\max iter \\ 1+10^3\ Iter/\max iter & if \quad 0.5\max iter < Iter < 0.75\max iter \\ 1+10^2\ Iter/\max iter & if \quad 0.1\max iter < Iter < 0.50\max iter \\ 1 & otherwise \end{cases} \quad (12)$$

Algorithm 1 Pre-processing using MALO

Input :Number of ant lions, fitness function, Number of ants, max iteration
Output : Ant lion elite level and its value of fitness

1 Initialize ants and ant lions positions randomly
2 Compute the ant lion fitness value
3 Find the best ant lion

$$N_{oB} = \begin{bmatrix} g([B_{1,1,}\ B_{1,2},\dots,B_{1,c}]) \\ g([B_{2,1,}\ B_{2,2},\dots,B_{2,c}]) \\ \vdots \\ \vdots \\ \vdots \\ g([B_{m,1,}\ B_{m,2},\dots,B_{m,c}]) \end{bmatrix}$$

4 If suspension did not meet the standard

Volume 11, Issue 4
August 2022

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

5 Choose ant lion by using a wheel and trap
$d_j^T = antlion_j^T + d^T$

6 Produce a walk for each ant randomly

7 Update the position of each ant

8 Replace the ant with its fit

9 Upgrade Elite ant lion using
$antlion = ant_j^T \quad if \quad (ant_j^T) < antlion_j^T$

10 End while

This is called the best solution (ant lion level) S_H^T is a well-preserved and excellent ant lion affects the movement of all ants. S_B^T is the blinking lion that you select is the wheel let, in the following equation

$$ant_j^T = \frac{S_B^T + S_H^T}{2} \tag{13}$$

Work out the last stage of the hunt when the ants are attracted to the cell and eat. Then, the ant lion will renew its position in next equation.

$$antlion = ant_j^T \quad if \quad (ant_j^T) < antlion_j^T \tag{14}$$

The working function of pre-processing using MALO was described in algorithm 1.

4.2 Fitness-Distance Balance (FDB) Based Coyote Optimization Algorithm:

Feature selection is the process of reducing the number of input variables when creating a prediction model. It is advisable to reduce the number of input variables to reduce modelling costs and in some cases improve model performance. This method uses the FDB-based coyote optimization (FDB-CO) algorithm for optimal feature selection which computes the best and optimal features among multiple features in the document. In the first stage of the FDB method, the distance of the solution candidates from the best solution is calculated. Some distance measurements include Minkowski (MI), Euclidean (EU) and Manhattan (MA), which can be used for remote calculations. Eq and EU metric calculates the remote values of each member q_{best} are calculated as given in Eq. The solution is in distance e vector d_q formed for the applicant

$$\forall_{j=1}^{m}, q_j \neq q_{best}, d_{q_j} = \sqrt{(y_{1q_j} - y_{1qbest})^2 + (y_{2q_j} - y_{2qbest})^2 + ... + (y_{mq_j} - y_{mqbest})}^2 \tag{15}$$

$$d_q = \begin{bmatrix} D_1 \\ \vdots \\ D_n \end{bmatrix}_{n\times 1} \tag{16}$$

In the second phase of the FDB system, the points for eliminating candidates are calculated. Normal fitness values $(normf)$ and candidates' average distance values $(normd_q)$ are used to calculate the score. The reason for using the default number scale is to avoid the dominance of these two parameters when calculating points. The weight coefficient (z) is used to calculate limb scores, which determines the effect of exercise and the values of distance. Based on this description and 0 < z < 1, the score of each candidate in the population is calculated. In this study we get w 0.5. The FDB sample method can be used to calculate the score.

$$\forall_{j=1}^{m}, q_j, R_{FDB^1 q_j} = z * norm\ f_{q_j} + (1-z) * norm\, d_{q_j} \tag{17}$$

$$\forall_{j=1}^{m}, q_j, R_{FDB^2 q_j} = norm\ f_{q_j} * norm\, d_{q_j} \tag{18}$$

The score vector (R_q) of the population as follows:

$$R_q \equiv \begin{bmatrix} R_1 \\ \vdots \\ R_n \end{bmatrix}_{n\times 1} \tag{19}$$

The inhabitants contains M* packs, and every pack contain M_C coyotes, $M_C \in$ M* coyotes. A shared setting of every coyote, soc', represent the choice variables $\bar{Y}$ of optimization criteria with D-location. At instant of time T^{th}, social condition 'soc', of b^{th} coyote within q^{th} pack $soc_b^{q,T}$ is represent as

$$soc_b^{q,T} = \bar{Y} \equiv (y_1, y_2, y_3 ... y_d) \tag{20}$$

The random beginning of the social situation of the coyote: The randomness of the social status of the coyote b^{th} belong to q^{th} at instantaneous T^{th} and i^{th} dimension are started in as

$$soc_b^{q,T} = WC_i + R_i \times (VC_i - WC_i) \tag{21}$$

In i^{th} of the search space D, WC_i and VC_i are lower limits of the decision variable and upper limits of the decision

variable respectively. R_i is referred as random number in the range [0,1]. Call the fitness process the initial solution $fit_b^{q,T}$ is estimated at the first conditions $soc_b^{q,T}$ according to

$$fit_b^{q,T} = F(soc_b^{q,T}) \tag{22}$$

The nature of intelligence, communication, and good coyote structure contribute to the enhancement of their conditions. The Coyote renewal process depends on two factors: "the best social conditions in the $alpha^{q,T}$, and the team's cultural tendencies $cult^{q,T}$. This behavior can be expressed in the following ways:

$$\delta_1 = alpha^{q,T} - soc_{cr1}^{q,T} \tag{23}$$

$$\delta_1 = cult^{q,T} - soc_{cr2}^{q,T} \tag{24}$$

where R_1 and R_2 are two random coyotes in q[th] pack, while the random control variables of the problem are i_1 and i_2.

Algorithm 2 FDB based Coyote optimization algorithm

Input : FDB parameters, $soc_b^{q,T}$

Output: δ_1

1. Begin
2. Select a distance metric
3. $\forall_{j=1}^{m}, q_j \neq q_{best}, d_{q_j}$
4. Normalize distance and fitness vectors within the range [0, 1]
5. search process based on their scores (Rq)
6. $R_q \equiv \begin{bmatrix} R_1 \\ \vdots \\ R_n \end{bmatrix}_{n\times 1}$
7. Compute the birth and death process of coyote.
8. $soc_b^{q,T} = WC_i + R_i \times (VC_i - WC_i)$
9. Select the solution candidates to be included in the
10. $q_a = (1 - q_t)/2$
11. End

The probability of scattering and contact is given qt and qa are specified by (27)-(28). R_i is stated as random number results which are located within the range of the variable of the i[th] quantity. $ranD_i$ is a random number in [0, 1].

$$q_t = \frac{1}{D} \tag{25}$$

$$q_a = (1 - q_t)/2 \tag{26}$$

The algorithm 2 shows the working function of the FDB based Coyote optimization algorithm.

4.3 Capuchin Search Algorithm Based Neural Network

Group documents are compared with different methods of compiling documents to find the best way to get diploma documents. The number of groups to obtain the required parameter, a clustering algorithm is used to solve an unknown data problem. The best group method is determined using the clustering process by analyzing the results. CSNN algorithm is used for optimal document clustering which improves the performance of clustering. Cappuccino's behavioral strategies in the environment are related to their effective ability to search for food sources on trees, terrain, and riverbanks. Then, developing a mathematical model with global and local search strategies is described as follows. Cappuccinos usually jump from tree to tree, walk long distances in the trees in search of food, or fall to the ground in search of food along the river banks. Therefore, the jump system is similar to Global Search. The third operating rule can be used to represent Y.

$$y = y_0 + u_0 T + \frac{1}{2} b T^2 \tag{27}$$

Where y denotes capuchin new position, y_0 indicates initial position, u_0 denotes initial velocity, and acceleration is referred as b and time instance is denoted as T.

The Capuchin, U, using the first law of motion can fix speed in a jump.

$$u = u_0 + bT \tag{28}$$

In the operating model, the y and x factors of the initial velocity of the capuchin can be determined.

$$\begin{aligned} u_{0y} &= u_0 \cos(\theta_0) \\ u_{0x} &= u_0 \sin(\theta_0) \end{aligned} \tag{29}$$

In the y and x direction, the initial velocities are represented as u_{0y} *and* u_{0x} respectively. In the positive y-direction the angle measured is denoted as H_0. The horizontal velocities u_y can be obtained using the equation 2 and 3.

$$\begin{aligned} u_y &= u_{0y} + b_y T \\ &= u_0 \cos(\theta_0) \end{aligned} \tag{30}$$

Here, the horizontal acceleration is represented as b_y which is set to 0. The cappuccino distance, y, can be simplified as follows:

Volume 11, Issue 4
August 2022

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

$$y = y_0 + u_0 \cos(\theta_0) T \tag{31}$$

For vertical displacement, H is obtained as follows:

$$x = x_0 + u_0 \sin(\theta_0) T + \frac{1}{2} b_x T^2 \tag{32}$$

The capuchin initial position and vertical acceleration are indicated as x_0 and b_x respectively. Here, vertical acceleration means the acceleration of a fall due to gravity equal to G. The launch height x_0 is equal to height H which is used to find T by solving the equation 6.

$$T = 2u_0 \sin(\theta_0) / G \tag{33}$$

$\sin(2\theta_0) = 2\sin(\theta_0)\cos(\theta_0)$ is used for,

$$y = y_0 + u_0^2 \sin(2\theta_0) / G \tag{34}$$

Here capuchin new position is indicated as y, initial position is y_0, the initial velocity also denoted as u_0 and gravitational acceleration is referred as G and leaping angle is referred as H_0.

From information theory, cluster entropy uses the concept of entropy and measures the "integrity" of clusters. Lower entropy refers to a more homogeneous cluster and vice versa. When considering the results of a cluster experiment, $p(j,i)$ is a class label in the j file and the probability is given by cluster i. The entropy of the cluster is as follows:

$$e_i = -\sum_j p(j,i) \log_2 p(j,i) \tag{35}$$

The total entropy of a group of clusters is calculated as the sum of the entropies according to the cluster weight:

$$e = \sum_i \frac{m_i}{m} e_i \tag{36}$$

In cluster i, the number of documents is denoted as m_i and the total number of documents is represented as m.

Algorithm 3 Optimal document clustering using CSNN

Input : Capuchin position and velocity
Output : Total class entropy

1 Initialize the values for the input.
2 Apply Third operating rule in equation
$$y = y_0 + u_0 T + \frac{1}{2} b T^2$$
$$y = y_0 + u_0 T + \frac{1}{2} b T^2 .$$
3 Substitute First law of motion by
$u = u_0 + bT$
4 Obtain the horizontal velocities.
5 Simplify the cappuccino distance.
6 Calculate the entropy of the cluster
7 Apply conditional probability
8 Compute the Total class-entropy
9 End

For each cluster we calculate $p(j \mid i)$, the ability to assign a document to cluster i that is included in class 1. For a class j the entropy is given using these probabilities.

$$e_j^* = -\sum_i p(i \mid j) \log_2 p(i \mid j) \tag{37}$$

Instead of combined probability we use conditional probabilities because probability is generalized to total probabilities, i.e. $p(j \mid i)$. The Eq.6 is an entropic expression, while Eq.1 is not a true entropic expression, because $z_j p(j,i) \# 1$. Total class-entropy is calculated as the sum of "entropy" and is calculated by the probabilities of the class:

$$e^* = \sum_j \frac{m_j}{m} e_j^* \tag{38}$$

In cluster j, the number of documents is denoted as m_j. The algorithm 3 represents the working function of the capuchin search based neural network.

5. RESULTS AND DISCUSSION

To evaluate the performance of proposed automated efficient document clustering (AEDC) with the Reuters and the 20 Newsgroup database. The experimental setup, description of database and performance metrics is described as follows: The computer runs Windows 10, and has 2 GB of RAM and an Intel i3 core processor. A proposed AEDC method is implemented in the Windows operating system with the help of Spyder (Python 3.7). The presentation of the proposed AEDC system is comparable to current state-of-the-art equipment: WIPLSA, ICGT, CBICA and RMFO in terms of precision, accuracy, recall, F-measure and rand coding.

5.1 Accuracy Analysis

Table 2 summarizes the accuracy comparison of the proposed CSNN technique with varying chunk sizes as 2, 4, 6, 8 and 10 and compared with existing state-of-art WIPLSA, ICGT, CBICA and RMFO techniques. Fig. 2a shows the accuracy comparison of proposed and existing techniques for the Reuter database. It clearly depicts the accuracy of the

Volume 11, Issue 4
August 2022

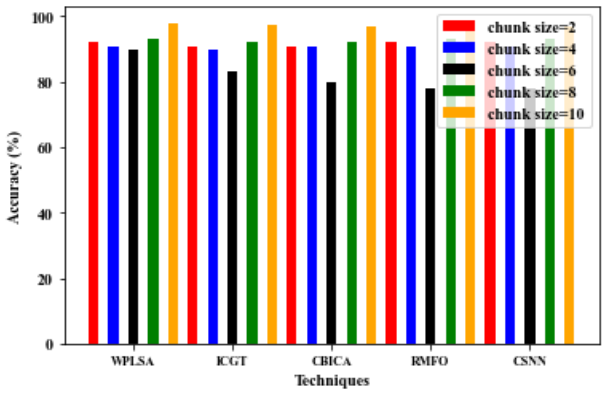
ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

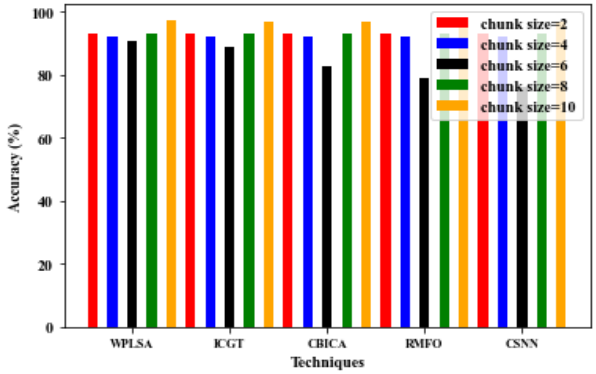
proposed CSNN technique is 4.7%, 6.6%, 15.8% and 4.7% higher than existing WPLSA, ICGT, CBICA and RMFO techniques respectively. Fig. 2b shows the accuracy comparison of proposed and existing techniques for 20 Newsgroups databases. It clearly depicts the accuracy of the proposed CSNN technique is 13.4%, 5.14%, 13.8% and 13.4% higher than existing WPLSA, ICGT, CBICA and RMFO techniques respectively.

Table 2 **Accuracy Comparison Of Proposed And Existing Techniques**

Techniques	Reuter database					20 Newsgroups database				
	2	4	6	8	10	2	4	6	8	10
WPLSA	92	91	91	92	92	93	93	93	93	93
ICGT	91	90	91	91	91	92	92	92	92	92
CBICA	90	83	80	78	78	91	89	83	79	76
RMFO	93	92	92	93	93	93	93	93	93	93
CSNN	98	97.5	97	97	96.8	97.5	97	97	96.8	97



(a)



(b)

Fig. 2 Accuracy Comparison Of Proposed And Existing Techniques (A) Reuter Database (B) 20 Newsgroups Database

5.2 Precession Analysis

Table 3 summarizes the precession comparison of the proposed CSNN technique with varying chunk sizes as 2, 4, 6, 8 and 10 and compared with existing state-of-art WIPLSA, ICGT, CBICA and RMFO techniques. Fig. 3a shows the precession comparison of proposed and existing techniques for the Reuter database. It depicts the precession of the proposed CSNN technique is 4.9%, 37.2%, 37.28% and 3.8% higher than existing WIPLSA, ICGT, CBICA and RMFO techniques respectively. Fig. 3b shows the precession comparison of the proposed and existing techniques for 20 Newsgroups databases. It depicts the precession of the proposed CSNN technique is 7.34%, 37.9%, 37.9% and 4.02% higher than existing WIPLSA, ICGT, CBICA and RMFO techniques respectively.

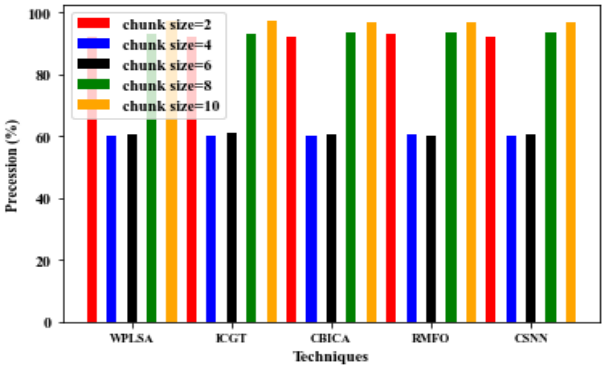
Table 3 Precession Comparison of Proposed and Existing Techniques

Techniques	Reuter database					20 Newsgroups database				
	2	4	6	8	10	2	4	6	8	10
WPLSA	92.1	92.3	92.1	92.9	92.1	92	92	92	87	85
ICGT	60.4	60.4	60.2	60.7	60.3	60	60	60	60	60
CBICA	60.8	60.9	60.7	60.3	60.7	60	60	60	60	60
RMFO	93.2	93.1	93.5	93.6	93.7	93	93	93	93	93
CSNN	97.5	97.3	97.1	97	96.7	97.3	97.1	97	96.7	96.5

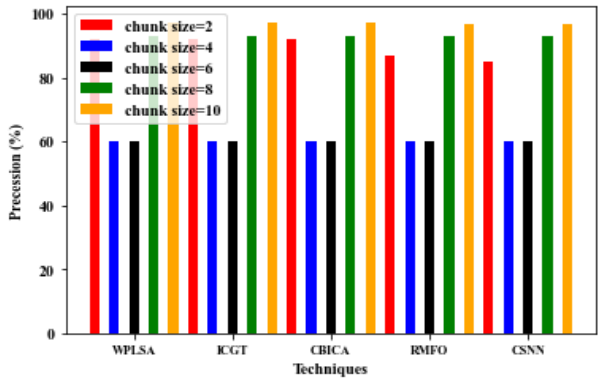
Volume 11, Issue 4
August 2022

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X



(a)



(b)

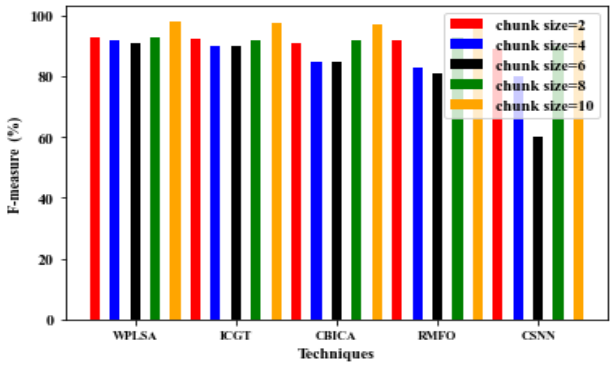
Fig. 3 Precession Comparison of Proposed and Existing Techniques (A) Reuter Database (B) 20 Newsgroups Database

5.3 F-measure analysis

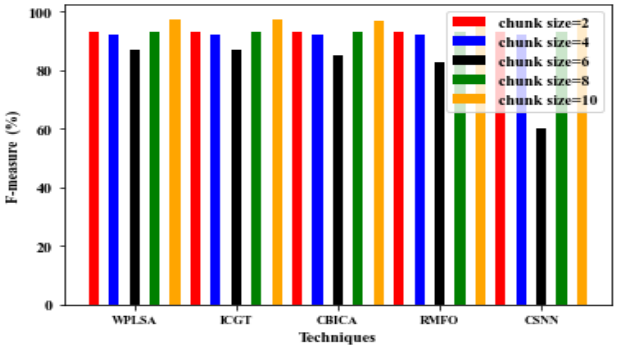
Table 4 summarizes the F-measure comparison of the proposed CSNN technique with varying chunk sizes as 2, 4, 6, 8 and 10 and compared with existing state-of-art WIPLSA, ICGT, CBICA and RMFO techniques. Fig. 4a shows the F-measure comparison of the proposed and existing techniques for the Reuter database. It depicts the F-measure of the proposed CSNN technique is 6.02%, 11.7%, 16.2% and 5.56% higher than existing WIPLSA, ICGT, CBICA and RMFO techniques respectively. Fig. 4b shows the F-measure comparison of the proposed and existing techniques for 20 Newsgroups databases. It depicts the F-measure of the proposed CSNN technique is 4.32%, 6.76%, 17.24% and 4.3% higher than existing WIPLSA, ICGT, CBICA and RMFO techniques respectively.

Table 4 **F-Measure Comparison of Proposed and Existing Techniques**

Techniques	Reuter database					20 Newsgroups database				
	2	4	6	8	10	2	4	6	8	10
WPLSA	93	92.5	91	92	89	93	93	93	93	93
ICGT	92	90	85	83	80	92	92	92	92	92
CBICA	91	90	85	81	60	87	87	85	83	60
RMFO	93	92	92	93	90	93	93	93	93	93
CSNN	98.2	97.5	97.3	97.1	97	97.5	97.3	97.1	97	97.2



(a)



(b)

Fig. 4 F-Measure Comparison of Proposed and Existing Techniques (A) Reuter Database (B) 20 Newsgroups Database

5.4 Recall Analysis

Table 5 summarizes the Recall comparison of the proposed CSNN technique with varying chunk sizes as 2, 4, 6, 8 and 10 and compared with existing state-of-art WIPLSA, ICGT, CBICA and RMFO techniques. Fig. 5a shows the Recall comparison of the proposed and existing techniques for the Reuter database. It depicts the Recall of the proposed CSNN technique is 3.4%, 4.5%, 11.8% and 2.6% higher than existing WIPLSA, ICGT, CBICA and RMFO techniques respectively. Fig. 5b shows the Recall comparison of the proposed and existing techniques for 20 Newsgroups

Volume 11, Issue 4
August 2022

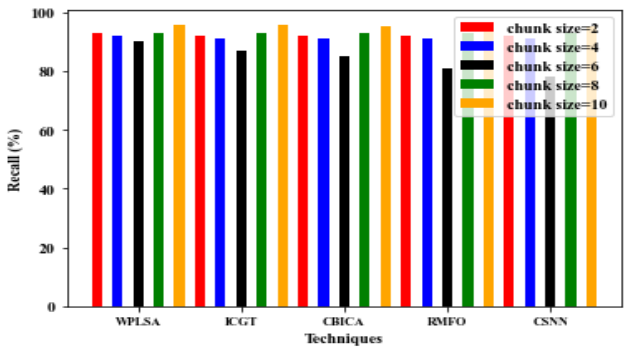
ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

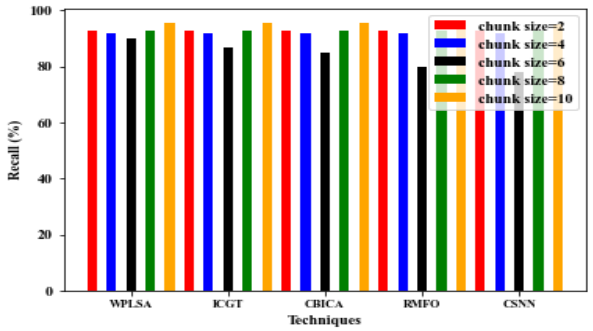
databases. It depicts the Recall of the proposed CSNN technique is 6.6%, 3.5%, 12.6% and 6.6% higher than existing WIPLSA, ICGT, CBICA and RMFO techniques respectively.

Table 5 **Recall Comparison of Proposed and Existing Techniques**

Techniques	Reuter database					20 Newsgroups database				
	2	4	6	8	10	2	4	6	8	10
WPLSA	93	92	92	92	92	93	93	93	93	93
ICGT	92	91	91	91	91	92	92	92	92	92
CBICA	90	87	85	81	78	90	87	85	80	78
RMFO	93	93	93	93	93	93	93	93	93	93
CSNN	95.9	95.7	95.5	95.4	95.3	95.7	95.5	95.4	95.3	95.2



(a)



(b)

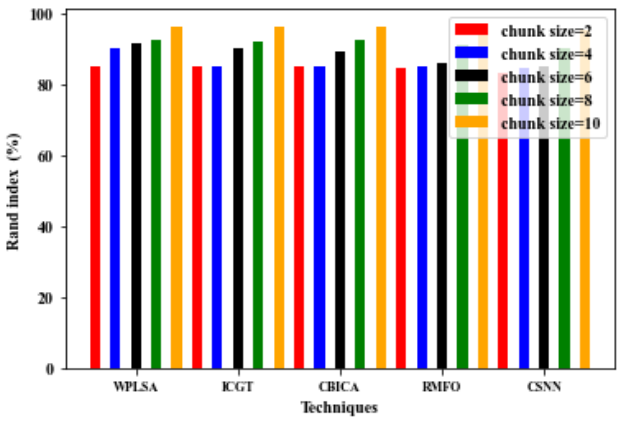
Fig. 5 Recall Comparison Of Proposed And Existing Techniques (A) Reuter Database (B) 20 Newsgroups Database

5.5 Rand Index Analysis

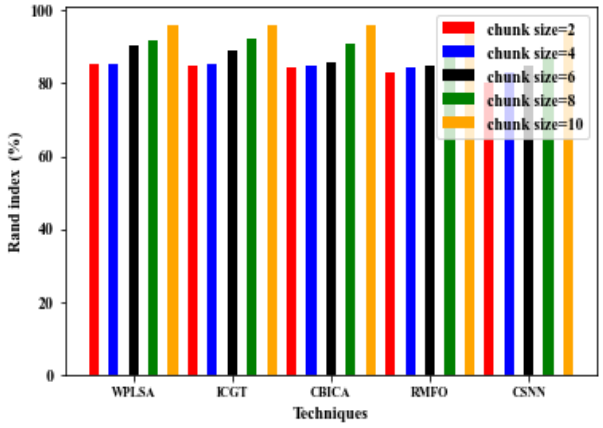
Table 6 summarizes the Rand index comparison of the proposed CSNN technique with varying chunk sizes as 2, 4, 6, 8 and 10 and compared with existing state-of-art WIPLSA, ICGT, CBICA and RMFO techniques. Fig. 6a shows the Rand index comparison of the proposed and existing techniques for the Reuter database. It depicts the Rand index of the proposed CSNN technique is 11.7%, 10.2%, 7.8% and 4.8% higher than existing WIPLSA, ICGT, CBICA and RMFO techniques respectively. Fig. 6b shows the Rand index comparison of the proposed and existing techniques for 20 Newsgroups databases. It depicts the Rand index of the proposed CSNN technique is 12.5%, 11.54%, 8.96% and 5.4% higher than existing WIPLSA, ICGT, CBICA and RMFO techniques respectively.

Table 6 **Rand Index Comparison of Proposed and Existing Techniques**

Techniques	Reuter database					20 Newsgroups database				
	2	4	6	8	10	2	4	6	8	10
WPLSA	85.2	85.1	85.0	84.5	83.2	85.1	85.0	84.5	83.2	80.2
ICGT	90.31	85.2	85.1	85.0	84.5	85.2	85.1	85.0	84.5	83.1
CBICA	91.4	90.23	89.2	85.8	85.0	90.2	89.2	85.8	85.0	84.8
RMFO	92.4	92.	92.3	90.9	90.0	92.0	92.3	90.9	90.0	87.3
CSNN	96.3	96.1	96.1	95.8	95.2	96.1	96.1	95.8	95.2	94.8



(a)



(b)

Fig. 6 Rand Index Comparison of Proposed and Existing Techniques (A) Reuter Database, (B) 20 Newsgroups Database

6. CONCLUSION

An automated efficient document clustering method is proposed using hybrid FDB-based coyote optimization and capuchin search-based neural network technique (AEDC). A modified ant lion optimization (MALO) algorithm is proposed for data pre-processing which removes the unwanted artifacts and redundant content from the documents. FDB-based coyote optimization (FDB-CO) algorithm is used for optimal feature selection which computes the best and optimal features among multiple features in documents. A capuchin search-based neural network (CSNN) is used for optimal document clustering which improves the performance of clustering. Finally, the proposed method evaluates using two datasets, namely the Reuter database, 20 Newsgroups database and own dataset. From the comparative analysis, it is observe that the average accuracy of the proposed AEDC method is 12.39% higher than current methods; the average precession of the proposed AEDC method is 11.90% higher than existing methods; the average F-measure of the proposed AEDC method is 8.14% higher than current methods; the average recall of the proposed AEDC method is 9.387% higher than existing methods; and the average rand index of the proposed AEDC method is 15.298% higher than existing methods.

REFERENCES

[1] Mei, J.P., Wang, Y., Chen, L. and Miao, C., 2016. Large scale document categorization with fuzzy clustering. *IEEE Transactions on Fuzzy Systems*, *25*(5), pp.1239-1251.

[2] Ling, Y., Pan, X., Li, G. and Hu, X., 2015. Clinical documents clustering based on medication/symptom names using multi-view nonnegative matrix factorization. *IEEE transactions on nanobioscience*, *14*(5), pp.500-504.

[3] Fu, X., Huang, K., Yang, B., Ma, W.K. and Sidiropoulos, N.D., 2016. Robust volume minimization-based matrix factorization for remote sensing and document clustering. *IEEE Transactions on Signal Processing*, *64*(23), pp.6254-6268.

[4] Chiang, I.J., Liu, C.C.H., Tsai, Y.H. and Kumar, A., 2015. Discovering latent semantics in web documents using fuzzy clustering. *IEEE Transactions on Fuzzy Systems*, *23*(6), pp.2122-2134.

[5] Gu, J., Feng, W., Zeng, J., Mamitsuka, H. and Zhu, S., 2012. Efficient semisupervised MEDLINE document clustering with MeSH-semantic and global-content constraints. *IEEE transactions on cybernetics*, *43*(4), pp.1265-1276.

[6] Spanakis, G., Siolas, G. and Stafylopatis, A., 2012. Exploiting Wikipedia knowledge for conceptual hierarchical clustering of documents. *The Computer Journal*, *55*(3), pp.299-312.

[7] Huang, R., Yu, G., Wang, Z., Zhang, J. and Shi, L., 2012. Dirichlet process mixture model for document clustering with feature partition. *IEEE Transactions on knowledge and data engineering*, *25*(8), pp.1748-1759.

[8] Zhang, X., Hu, X., Hu, T., Park, E.K. and Zhou, X., 2012. Utilizing different link types to enhance document clustering based on Markov Random Field model with relaxation labeling. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *42*(5), pp.1167-1182.

[9] Yan, L., Tang, W., Wu, Q.H. and Smith, J.S., 2017. Kernel-based consensus clustering for ontology-embedded document repository of power substations. *CSEE Journal of Power and Energy Systems*, *3*(2), pp.212-221.

[10] Zhang, T., Tang, Y.Y., Fang, B. and Xiang, Y., 2011. Document clustering in correlation similarity measure space. *IEEE Transactions on Knowledge and Data Engineering*, *24*(6), pp.1002-1013.

[11] Ling, Y., Pan, X., Li, G. and Hu, X., 2015. Clinical documents clustering based on medication/symptom names using multi-view nonnegative matrix factorization. *IEEE transactions on nanobioscience*, *14*(5), pp.500-504.

[12] Pei, X., Chen, C. and Gong, W., 2016. Concept factorization with adaptive neighbors for document clustering. *IEEE transactions on neural networks and learning systems*, *29*(2), pp.343-352.

[13] da Cruz Nassif, L.F. and Hruschka, E.R., 2012. Document clustering for forensic analysis: An approach for improving computer inspection. *IEEE transactions on information forensics and security*, *8*(1), pp.46-54.

[14] Abualigah, L.M., Khader, A.T. and Hanandeh, E.S., 2018. A combination of objective functions and hybrid krill herd algorithm for text document clustering analysis. *Engineering Applications of Artificial Intelligence*, *73*, pp.111-125.

[15] Sardar, T.H. and Ansari, Z., 2018. An analysis of MapReduce efficiency in document clustering using parallel K-means algorithm. *Future Computing and Informatics Journal*, *3*(2), pp.200-209.

[16] He, S., Samara, P., Burgers, J. and Schomaker, L., 2016. A multiple-label guided clustering algorithm for historical document dating and localization. *IEEE Transactions on Image Processing*, *25*(11), pp.5252-5265.

[17] Yan, W., Zhang, B., Ma, S. and Yang, Z., 2017. A novel regularized concept factorization for document clustering. *Knowledge-Based Systems*, *135*, pp.147-158.

[18] Huang, S., Xu, Z. and Lv, J., 2018. Adaptive local structure learning for document co-clustering. *Knowledge-Based Systems*, *148*, pp.74-84.

[19] Ma, Y., Wang, Y. and Jin, B., 2014. A three-phase approach to document clustering based on topic significance degree. *Expert systems with applications*, *41*(18), pp.8203-8210.

[20] Laclau, C. and Nadif, M., 2016. Hard and fuzzy diagonal co-clustering for document-term partitioning. *Neurocomputing*, *193*, pp.133-147.

[21] Sheri, A.M., Rafique, M.A., Hassan, M.T., Junejo, K.N. and Jeon, M., 2019. Boosting discrimination information based document clustering using consensus and classification. *IEEE Access*, *7*, pp.78954-78962.

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

[23] Yarlagadda, M., Rao, K.G. and Srikrishna, A., 2019. Frequent itemset-based feature selection and Rider Moth Search Algorithm for document clustering. *Journal of King Saud University-Computer and Information Sciences*.

[24] Bikku, T. and Paturi, R., 2019. A novel somatic cancer gene-based biomedical document feature ranking and clustering model. *Informatics in Medicine Unlocked*, *16*, p.100188.

[25] Curiskis, S.A., Drake, B., Osborn, T.R. and Kennedy, P.J., 2020. An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. *Information Processing & Management*, *57*(2), p.102034.

[26] Abasi, A.K., Khader, A.T., Al-Betar, M.A., Naim, S., Makhadmeh, S.N. and Alyasseri, Z.A.A., 2020. Link-based multi-verse optimizer for text documents clustering. *Applied Soft Computing*, *87*, p.106002.

[27] Park, J., Park, C., Kim, J., Cho, M. and Park, S., 2019. ADC: Advanced document clustering using contextualized representations. *Expert Systems with Applications*, *137*, pp.157-166.

[28] AlMahmoud, R.H., Hammo, B. and Faris, H., 2020. A modified bond energy algorithm with fuzzy merging and its application to Arabic text document clustering. *Expert Systems with Applications*, *159*, p.113598.

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

[30] Kim, H., Kim, H.K. and Cho, S., 2020. Improving spherical k-means for document clustering: Fast initialization, sparse centroid projection, and efficient cluster labeling. *Expert Systems with Applications*, *150*, p.113288.

[31] Yarlagadda, M., Kancherla, G.R. and Atluri, S., 2019. Incremental document clustering using fuzzy-based optimization strategy. Evolutionary Intelligence, pp.1-14.

AUTHOR PROFILES

Dr. P. Perumal, received his Ph.D in Computer Science and Engineering from Anna University, Chennai. He holds his Masters in Computer Applications from Bharathidasan University, Software Engineering from Anna University and Bachelors in Mathematics from Madras University. He has about 50 technical and research publications and presentations to his credit in International and National journals and conferences. His research areas include Data Mining and Image Processing. He has 25 years of academic experience and one year industry experience. He is the SBC of CSI Student Branch in SREC. He is a reviewer in leading journals like Knowledge based systems (Elsevier), Knowledge Information System, Journal of Information Technology, Journal Computers, IET Image Processing and WSEAS. He is a life member of ISTE and member of various professional societies like CSI, IAENG, IACSIT and IDES. He has received an Academic Leadership Award in the year 2020, IARDO Award for Excellence in 2018 and Longest SBC Award from CSI (4 times).