Resbee Publishers

Multimedia Research (MR)

Received 29 August, Revised 3 October, Accepted 19 October

A Semantic Word Processing Using Enhanced Cat Swarm Optimization Algorithm for Automatic Text Clustering

Vidyadhari Ch.
St. Joseph College of Engineering
Chennai, Tamilnadu, India
vidyadhari.ch01@gmail.com

Sandhya N.
VNRVJIET
Hyderabad, Telangana, India

Premchand P.
University College of Engineering, Osmania University,
Hyderabad, Telangana, India

Abstract: Generally, Text mining indicates the process of extracting maximum-quality information from the text. Moreover, it is mostly exploited in applications such as text categorization, text clustering, and text classification and so forth. In recent times, the text clustering is considered as the facilitating and challenging task exploited to cluster the text document. Because of the few inappropriate terms and large dimension, accuracy of text clustering is reduced. In this work, the semantic word processing and Enhanced CSO algorithm are presented for automatic text clustering. At first, input documents are stated as input to the preprocessing step that provides the useful keyword for clustering and feature extraction. After that, the ensuing keyword is applied to wordnet ontology to discover the hyponyms and synonyms of every keyword. Then, the frequency is determined for every keyword used to model the text feature library. Since it comprises the larger dimension, the entropy is exploited to choose the most significant feature. Hence, the proposed approach is exploited to assign the class labels to generate different clusters of text documents. The experimentation outcomes and performance is examined and compared with conventional algorithms such as ABC, GA, and PSO.

Keywords: text mining; clustering; text clustering; text documents; word processing; optimization algorithm

Nomenclature

Abbreviations	Descriptions
FAQ	Frequently Asked Questions
TF	Term Frequency
ACO	Ant Colony Optimization
JSD	Jensen Shannon-divergence
SI	Swarm Intelligence
VSM	Vector Space Model
ML	Machine Learning
IDF	Inverse Document Frequency
NLP	Natural Language Processing
SCPSO	Spectral Clustering algorithm with Particle Swarm Optimization
TRTD	Topic Representative Term Discovery
GA	Genetic Algorithm
SPC	Self Position Consideration
MOO	Multi-Objective Optimization
GSA	Gravitational Search Algorithm
MR	Mixture Ratio
ABC	Artificial Bee Colony
SMP	Seeking Memory Pool
PSO	Particle Swarm Optimization
CDC	Counts of Dimension to Change
CSO	Cat Swarm Optimization
SRD	Selected Dimension

1. Introduction

In recent days, the clustering approach is utilized in surfing documents collections as well as to normalize the outcomes that are stated using the search engine on the basis of the user's query. Generally, text clustering is exploited to extract the relevant features and indicates the features in significant manners [1]. In-text mining system, the documents are indicated as maximum dimensional documents with complex semantics. Common applications of document clustering are document organization, automatic topic extraction, and information retrieval [2]. Nevertheless, numerous significant research works were performed in the topic of text clustering that requires the determination to enhance and improve the text clustering process quality.

A growing number of short texts were produced such as forum titles, search result snippets, image or video tags and titles, FAQ, microblogs, tweets, and hitherto in the present Web 2.0 epoch. This has followed in an increasing required for speed and effectual clustering of short texts based on maximum similitude and dissimilarities among clusters. A well-planned brief text clustering algorithm can significantly stimulate and encourage its real applications like the image or video tagging, topic detection, answering service recommendations, information retrieval, so forth. However, different from usual texts, exploit of brief texts is complex using sparsity as well as maximum dimensionality. Hence, the traditional tf-idf measure, the VSM, and usual text clustering algorithm might not work effectively while utilized to brief texts [4].

Text clustering requires partitioning the disjoint subsets from the source cluster of texts hence each cluster must contain the same objects, in addition to the different clusters object vary notably between themselves [16] [17]. Clustering is referred as an unsupervised learning algorithm. The outcome (the partition, the clustering,) is on the basis of only on the similarity measure, object representation, and the clustering algorithm. If these agree almost equally to the user's comprehension the outcome might effectively be an instinctive and functional clustering. Consider, even though that clustering algorithms always create clusterings, while this is not stated, and that therein major scenarios subsist numerous appropriate clusterings of a set of complex objects.

Various kinds of research were exploited to resolve this chore by creating the issue into an optimization chore [6]. Moreover, the issue is resolved using an ACO in integration with JSD as a fitness function [7]. Initially, to secure the ACO of probable cycling, the graph coloring module is exploited to establish various kinds of brief messages which possess different content [8]. For more information, [8] proposed various summarization algorithms on the basis of the SI approaches. Conversely, few algorithms depend on the ML approach in addition to feature extraction to design the précis chore as a suggested as well as classification task [9]. In addition, in the short texts, few hidden emotions were recognized on the basis of sentiment analysis [10]. Conversely, to improve the précis task on micro-blogs, various examinations were utilized to attain the summarization task [11]. Conversely, to obtain effects of product's features from user's standpoints, other algorithms were proposed on the basis of the NLP, in addition to sentiment analysis affiliated with ML [12].

The main contribution of this work is to exploit the WordNet Ontology to discover the semantic words of every document to model the feature library. In addition, due to the larger dimension of the feature library, the entropy is examined to choose the most significant feature to carry out the text clustering. The novel enhanced CSO algorithm is proposed it used for automatic text clustering to calculate the fitness function.

2 Literature Review

In 2019, R. Janani and Dr. S. Vijayarani [1] presented a novel SCPSO to enhance the text document clustering. The randomization was performed with the preliminary population by considering local and global optimization model. It aspires at integrating the spectral clustering with swarm optimization to pact with the enormous volume of text documents. In 2019, Shuiqiao Yang et al. [2] presented a novel TRTD algorithm for brief text clustering. In this TRTD algorithm, clusters of intimately bound up topic typical words were discovered using the nearness and importance of words. The nearness of the topic typical words was estimated using their interdependent co-occurrence, and the importance was evaluated by their global word incidences during the complete brief text corpus. In 2019, Mohamed Atef Mosa [3], developed mining in large social media data into a MOO chore for the primary time to extract the spirit of a term. As few users might insist the summary at any instant, various clusters of dissimilar brief messages were recognized on the basis of the graph coloring method. Moreover, a GSA was exploited to assure various significant objectives for producing a brief summary. In 2017, Elizaveta K.Mikhina and Vsevolod I.TRifalenkov [4], suggested a algorithm of text clustering, which does not depend on any user-set parameters. Text documents and relationships between them were indicated as graph edges and nodes and graph community recognition algorithm was hence developed to the text clustering issue. The algorithm was examined over news articles clusters and exhibited efficient automatic clustering of text documents in clusters were similar or actually close. In 2017, Caiyan Jia et al [5], presented a novel concept decomposition algorithm, which produces concept vectors using identifying semantic term communities. By mapping the novel brief texts to the learned semantic concept vectors were evaluated the cluster memberships for brief texts. The proposed algorithm was not only vigorous to the sparsity of brief text corpora however as well conquers the dimensionality curse, scaling to a great quantity of brief text inputs because the concept vectors being attained from word-word rather than document-word space.

3. Proposed Methodology

3.1 Automatic Text Clustering and Semantic Word Processing using Proposed Algorithm

The main aim of this article is to develop and model the novel clustering algorithm by exploiting the enhanced CSO algorithm and semantic word processing. For the proposed method, different number of documents is considered as the input. Schematic diagram of the proposed methodology is demonstrated in Fig. 1. Initially, the input documents are subjected to the preprocessing step whereas the useful keywords for the document can be recognized with the help of stop word stemming and removal algorithm. Subsequently, to extract the antonyms and synonyms of every keyword the wordnet ontology is exploited. These keywords are subsequently exploited to model the feature library on the basis of the semantic words and keywords frequency attained from the wordnet. Because of larger feature library dimension, and the entropy measure is utilized in order to choose the most significant feature. To resourcefully group the input document, new enhanced CSO is proposed. Subsequently, proposed algorithm uses the objective model of fuzzy metrics on the basis of the Kernel. At last, the kernel function is exploited to estimate the fitness model of the proposed algorithm whereas the text clustering mechanism is done considerably.

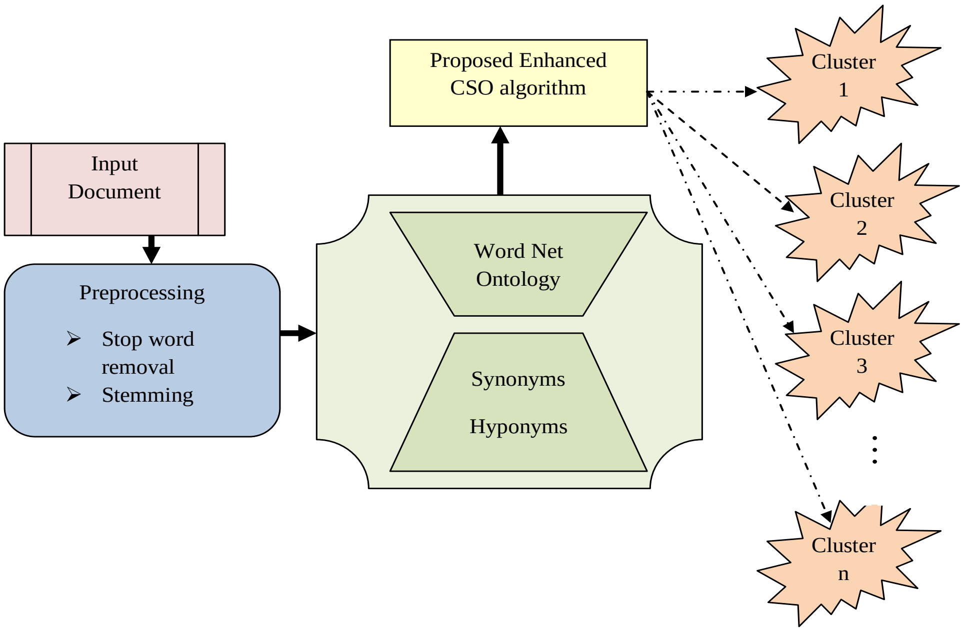


Fig. 1. *Block diagram of the proposed algorithm*

3.2 Preprocessing

At first, the input document undergoes for the preprocessing step. In the preprocessing step, the stemming algorithm and stop word removal are exploited. The motivation of preprocessing is to extract the notable information and in addition to minimizing the document dimension. This step provides the main significant terms for the text clustering. The explanation for the preprocessing step is stated as follows.

Initially, the stop text removal is exploited to the input document D. Generally, the document comprises a high count of stop texts such as 'who', 'that', 'a', 'for' 'an', 'those', 'which', so forth are evaded in the preprocessing step. As the stop text performs minimum significant information of the document, the removal of stop word is exploited to minimize deceive and worsen performance.

Subsequently, in this step, the stemming is exploited to minimize the insignificant words of the input document. The aim of the stemming approach is to transform the derivative or inflection terms of the document into the base root structure. Hence, the porter stemming algorithm [13] is generally exploited for the text clustering. Following to the preprocessing, the w numbers of functional keywords are attained that is subsequently subjected to the wordnet ontology.

3.3 Semantic Word Processing: WordNet

The conventional algorithms do not explain the document semantics to execute the text clustering. To tackle this problem, the wordnet provides the word frequency exploited to discover the hyponyms and synonyms of the word in the document. Hence, to the wordnet semantic processing, the resultant keywords are subjected. The document semantic is notably indicated by exploiting the wordnet ontology. Generally, Wordnet is considered a method of lexical database, which consists of both thesaurus as well as dictionary. Hence, the lexical definitions of wordnet are verbs, nouns, adverbs, and adjectives. WordNet is exploited whereas 117,597 senses, 155,327 terms, and 207,016 pairs are indulged. By significant keywords, the synsets are arranged that provides the hyponyms, synonyms of its equivalent word of the input document. The synsets are represented as one of the significant issues that are indicated by integrating both lexical relations and conceptual-semantic. Therefore, the WordNet is exploited to integrate the semantic features of the term content improves the precise of the clustering.

To WordNet ontology, the k counts of a keyword are exploited whereas the synonyms and hyponyms can find out. The significant WordNet ontology factor is to exploit the semantic word that creates the cluster to be comprehensible as well as natural. Fig 2 states the diagrammatic illustration of WordNet ontology.

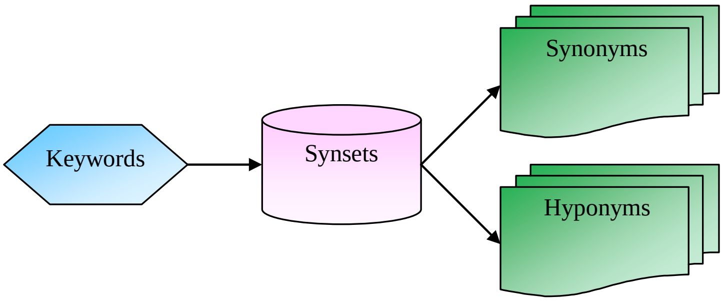


Fig. 2. *Diagrammatic illustration of WordNet ontology*

The synonyms provide the symmetrical connection among two corresponding and near ideas of the document. It is exploited to modified by another text without any alteration of its equivalent word.

The hyponym describes the connection between narrower and generic word. It is indicated in WordNet using the sign '@' that indicates "is-a" and "is a kind of".

In fig. 2, the w keywords are subjected to the wordnet ontology whereas it uses the synsets exploited to extract the hyponyms and synonyms, which is stated in eq. (1).

$$W_i = \{P_i, Q_i\}; \quad 1 \le i \le w \tag{1}$$

In eq. (1), i denotes the total number of keywords, P_i and Q_i denotes the synonym and hyponyms of i^{th} the keyword.

3.4 Feature Library on the Basis of the Entropy Measure

Once the hyponym and synonym of every keyword are determined, it is subsequently exploited to model the feature library. By exploiting the Wordnet Ontology, the w hyponym and synonym of w keywords are attained. Hence, the v numbers of notable keywords are attained that is subsequently exploited to model the feature library. Fig. 3 illustrates the representation of the feature matrix. The subsequent steps are described the feature selection on the basis of the entropy measure.

The feature selection is mostly exploited to minimize the input document dimension. Hence, prior to the clustering performance, the proposed algorithm needs the feature space of the document. The feature vector or space is decided on the basis of the keywords for the input document. Regarding the feature matrix, by calculating the frequency, the features are chosen for each v keyword in the dictionary learning.

Each keyword frequency is computed by how many times the keyword is exploited in its equivalent document in dictionary learning. Hence, the often exploited keywords are counted yields the frequency measure. For instance, w_{cd} is the d^{th} keyword of c^{th} the document. If the keyword b is occurring 50 times in document c, subsequently the frequency of the keyword w_{cd} is 50.

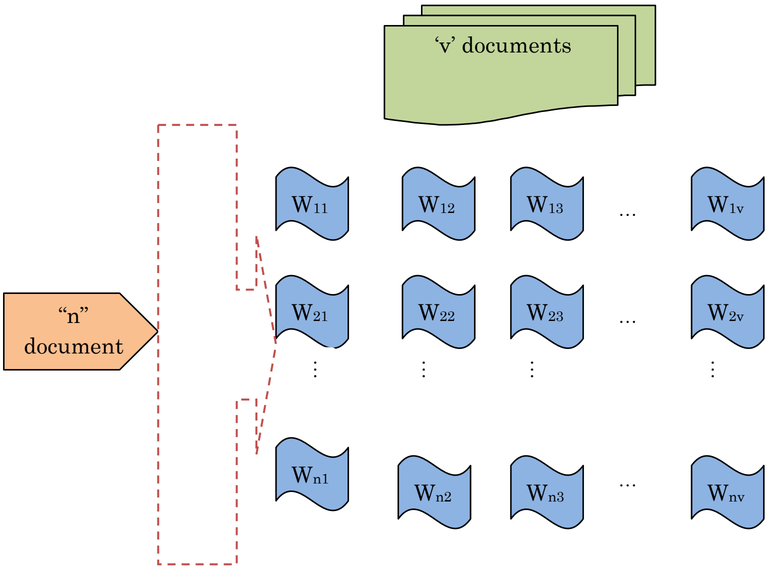


Fig. 3. *Diagrammatic representation of Feature Matrix*

Then, the document frequency is used to assign the frequency or score for each keyword. Moreover, it calculates the count of often exploited texts by the input document as it comprises the notable information. Each keyword exploited to construct the feature space on the basis of the value of the frequency and semantic terms such as hyponyms and synonyms attained from the wordnet. Hence, the f_f count of features is attained which is described by the feature matrix size $l \times v$ whereas in the dictionary learning v forms a high number of keywords.

3.4.1 Entropy Measure for Selection

Because of the larger dimension of the feature space, it produces the calculation burden to the performance of the clustering. In this paper, the entropy measure is utilized to solve this problem. Generally, the documents are allocated using the class labels at first. The keyword in each document is evaluated with the class label based on the class label, to calculate the entropy measure. Here, the entropy is described as the unpredictability of the information content and measure of energy. Moreover, the entropy is stated as the sum of the minimum probability measure of the features of each document. The entropy is exploited to the attained features whereas the minimum feature is used for the performance of clustering. Eq. (2) is used to determine the entropy.

$$E_R = -\min_{i \in 1,2,..l} \sum_{j=1}^{v} P_j \log(P_j) \qquad (2)$$

In eq. (2), P_j states the probability measure. Subsequently, the feature is chosen on the basis of the entropy value exploited to solve the dimensionality issue. Hence, feature library size attains and it minimized to $l \times u$; where $u < v$ whereas, l represents the count of documents and u describes the total number of features that are subjected to the proposed enhanced CSO algorithm.

4. Text Clustering using Proposed Enhanced Cat Swarm Optimization (CSO) Approach

4.1 Conventional CSO Approach

In CSO algorithm, the cats are the possible solution of the problem to be resolved. In the real world, the cat is modeled, a number of cats are on the circumstances of looking and lazy around named seeking mode, the remaining cat is tracking targets named tracking mode. The fundamental process of CSO algorithm [14] is stated below:

(a): Generate N cats as the swarm.

(b): Describe the parameter of SPC. Then arbitrarily partition the cats into seeking mode that put $SPC = 1$ as well as tracking mode that set $SPC = 0$ according MR.

(c): Use the cat into the seeking mode process while its SPC is one as well as use the cat into the tracking mode process while its SPC is 0. According to the process of (a) and (b) update the tracking mode and seeking mode.

(d): On the basis of the fitness model evaluates the cats and keeps the location of the optimal cat into memory.

(e): Verify the terminating state, if attained, terminate the or else repeat the (b) to (d).

The seeking mode is equivalent to a local search process for the optimization issue. The seeking mode has three significant parts such as the SMP that is the replicas of a cat created in the seeking mode, the SRD seeking range of that is the mutative portion for the chosen dimensions and CDC. The seeking mode process is stated below:

(a): If $SPC = 1$, $setT(= SMP)$ copies of cat_j.

(b): Based on CDC, use the mutation operator to the T copies. Arbitrarily subtracts or adds SRD ratio of the current values, restore the past values.

(c): The mutated copies fitness is evaluated.

(d): Utilizing eq. (3) computes the probability of selection for every candidate as well as chooses the point with maximum selecting probability to put back cat_j. If the objective of the fitness model is to discover the least amount solution, let $FF_a = FF_{min}$ else $FF_a = FF_{max}$

$$P_m = \frac{|FF_i - FF_a|}{FF_{max} - FF_{min}}, \text{where } 0 < m < n \tag{3}$$

The tracing mode is equivalent to a global search process of the optimization issue. The tracing mode process is stated as below.

(a): If $SPC = 0$, by the eq. (4) update the velocities of every dimension of a cat.

(b): Verify if the velocities are in the span of minimum and maximum velocities, if it is over-span, set it is equivalent to the limit.

(c): By the eq. (5), update the locations.

(d): Verify if the locations are in the span of minimum and maximum locations, if it is over-range, set it to be equivalent to the limit.

$$u_{j,c}(t+1) = u_{j,c}(t) + ct \times rn \times (y_{best,c}(t) - y_{j,c}(t)) \tag{4}$$

$$y_{j,c}(t+1) = y_{j,c}(t) + u_{j,c}(t+1) \tag{5}$$

In eq. (4), $u_{j,c}(t)$ signifies the velocity of cat_j at iteration t on the dimension, $y_{j,c}(t)$ denotes the location of the cat_j at iteration t on the dimension c, $c, y_{best,c}(t)$ signifies the location of the cat_j who possess the optimal fitness value at iteration t on dimension c, rn denotes a random value in the range of [0,1] and ct_1 denotes a constant. In general, PSO is unbalanced and simple to fall into local optimal solution.

4.2 Proposed Enhanced CSO Approach

To enhance the search accuracy of the CSO algorithm, this paper adopts the adaptive inertia weight $\omega(t)$ [15]. Using eq. (6), the enhancement of the proposed algorithm is stated.

$$u_{j,c}(t+1) = \omega(t) \times u_{j,c}(t) + ct \times rn \times (y_{best,c}(t) - y_{j,c}(t)) \tag{6}$$

In eq. (6), $\omega(t)$ represents the adaptive inertia weight, $\omega(t)$ value alters with the range in the cat location. In a two dimensional space, assume each cat's location after and prior to altering reflected in the diagram $y_{j,1}(t)$ and $y_{j,2}(t)$ are the location correlates of cat_j at iteration t, $y_{j,1}(t+1)$ and $y_{j,2}(t+1)$ are the location correlates of cat_j at iteration $t+1$, hence obtain the cat's disarticulation as inertia weight of cat_j at subsequent iteration.

$$\omega_j(t+1) = \sqrt{\begin{bmatrix} y_{j,1}(t+1) \\ - y_{j,1}(t) \end{bmatrix}^2 + \begin{bmatrix} y_{j,2}(t+1) \\ - y_{j,2}(t) \end{bmatrix}^2} \tag{7}$$

The eq. (7) can be modified to D dimensions; hence the eq. (7) can be extended as eq. (8). At last, obtain the average of N cats' inertia weight represented as the adaptive inertia weight in the subsequent iteration, like eq. (9).

$$\omega_j(t+1) = \sqrt{\begin{bmatrix} y_{j,1}(t+1) \\ - y_{j,1}(t) \end{bmatrix}^2 + \begin{bmatrix} y_{j,2}(t+1) \\ - y_{j,2}(t) \end{bmatrix}^2 + \ldots + \begin{bmatrix} y_{j,D}(t+1) \\ - y_{j,D}(t) \end{bmatrix}^2} \tag{8}$$

$$\omega(t) = \frac{\sum_{j=1}^{N} \omega_j(t)}{N} \tag{9}$$

In eq. (9), the search space is cat's location difference is great is stated, also ensuing in the equivalent inertia weight will be high that is helpful to a global search. In subsequent iterations, cat's locations alter minimum, inertia weight is respectively minimized that is contributing to comparatively big in previous iteration of the algorithm, hence to enhance the optimization precision. The inertia weight alters with the move of cats' location and pacts with the standard of large in previous phase and little in subsequent phase that is more conductive to the convergence of the algorithm. Fig 4 exhibits the flow chart of the proposed enhanced CSO algorithm.

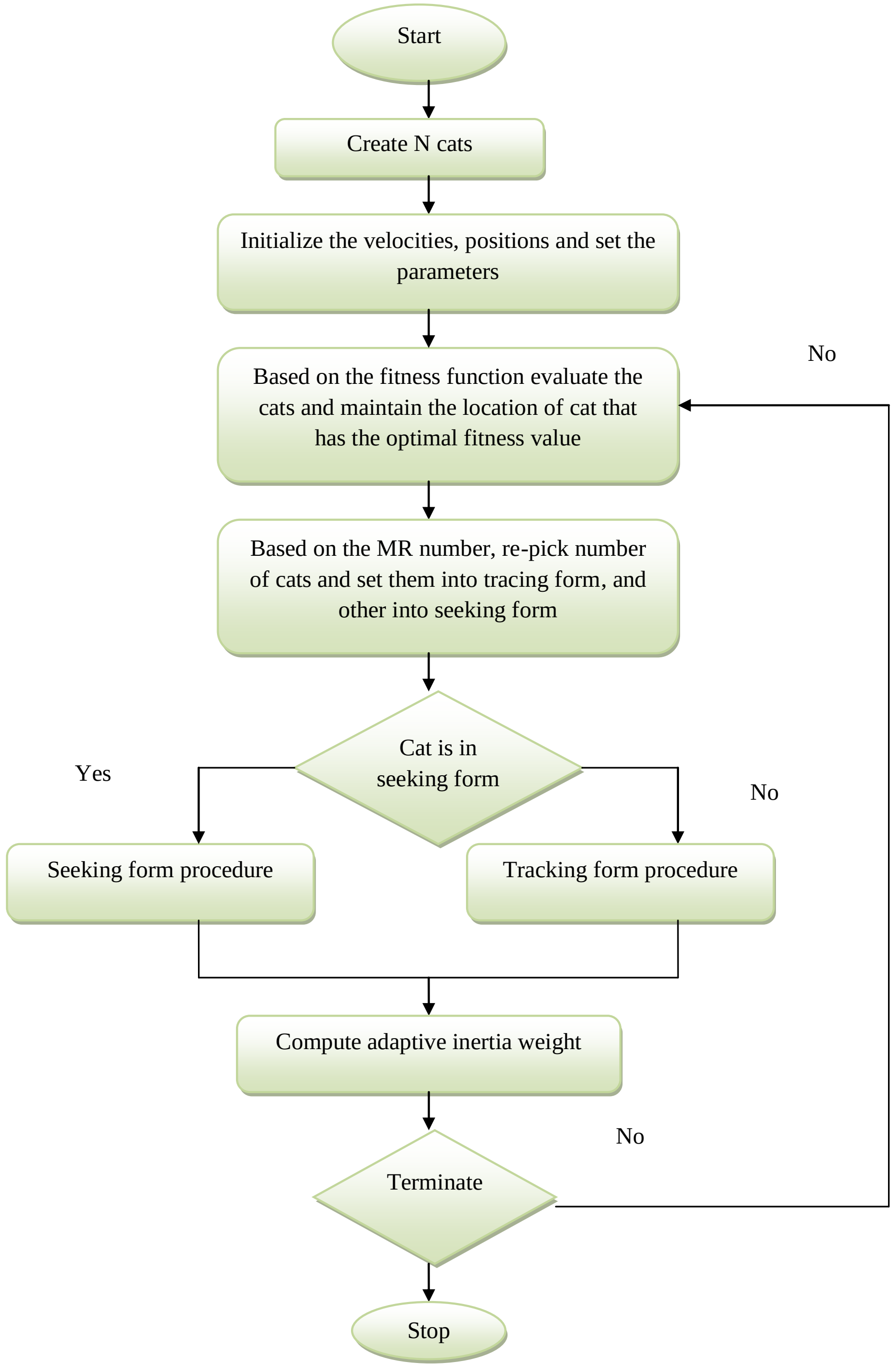


Fig. 4. *Flow chart of proposed Enhanced CSO algorithm*

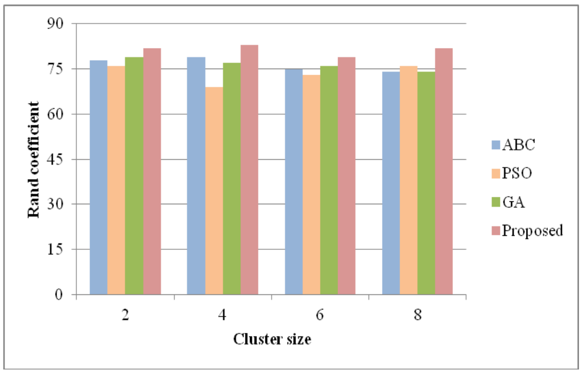
5. Results and Discussions

5.1 Simulation Procedure

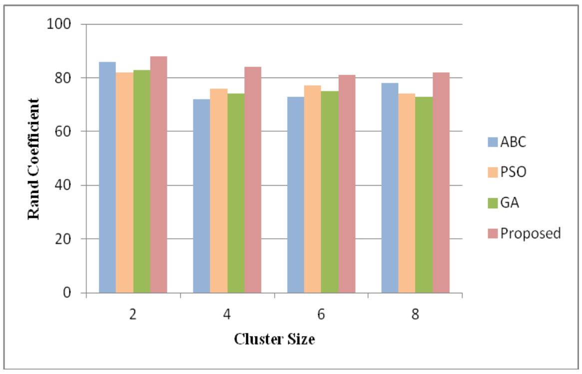
In the simulation of text clustering, the dataset such as 20 newsgroups [19] and Reuter [18] are exploited. The data in the Reuters is obtained by Carnegie group, Inc and Reuters, Ltd. pose with 21, 578 text documents. It becomes an efficient dataset for text clustering. Moreover, 20 newsgroups are a dataset described by the compilation of 20,000 newsgroups documents distinguished by the 20 several newsgroups. For evaluation, the Jaccard, clustering accuracy, and the rand coefficients are exploited.

5.2 Performance Evaluation

Fig. 5 shows the performance analysis of the proposed method and conventional methods for rand coefficients on the basis of the cluster size. Fig 5 (a) depicts the analysis exploiting 20 newsgroup data set. Here, the proposed method obtains maximum rand index value while comparing with the ABC, GA and PSO methods. Likewise, the performance evaluation of rand coefficient by Reuter is demonstrated in figure 5 (b). Comparing with the conventional methods, the proposed method obtains maximum rand index value. Hence, the fig 5 assures to enhance the similarity measure among two text clusters.



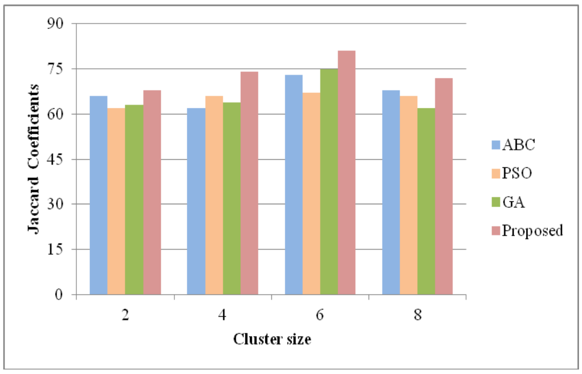
(a)



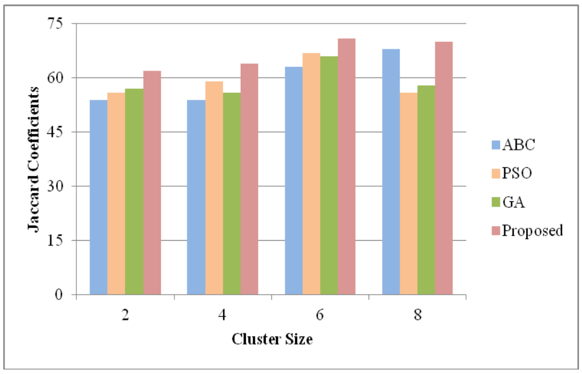
(b)

Fig. 5. *Performance analysis of Rand coefficient (a)20 Newsgroup; (b)Reuter*

In fig 6, the performance evaluation of the coefficient of the Jaccard is on the basis of the cluster size is exhibited. The coefficient of the Jaccard is the similarity measure of words among the clustered documents. In Fig 6 (a), the performance evaluation exploiting 20 newsgroup database is shown. Subsequently, the proposed algorithm compared with the conventional algorithms PSO, ABC, and GA, the proposed method is high which is somewhat maximized and it is exhibited in Fig. 6 (a). Subsequently, Fig 6 (b) demonstrates the performance evaluation exploiting Reuter database. Here, the proposed method attains better similarity of terms among two documents while comparing with the conventional algorithms. Hence, the performance of the developed text clustering is enhanced in that the significant information in a well-organized manner can be extracted.



(a)



(b)

Fig. 6. *Performance evaluation of Jaccard coefficient (a)20 Newsgroup; (b)Reuter*

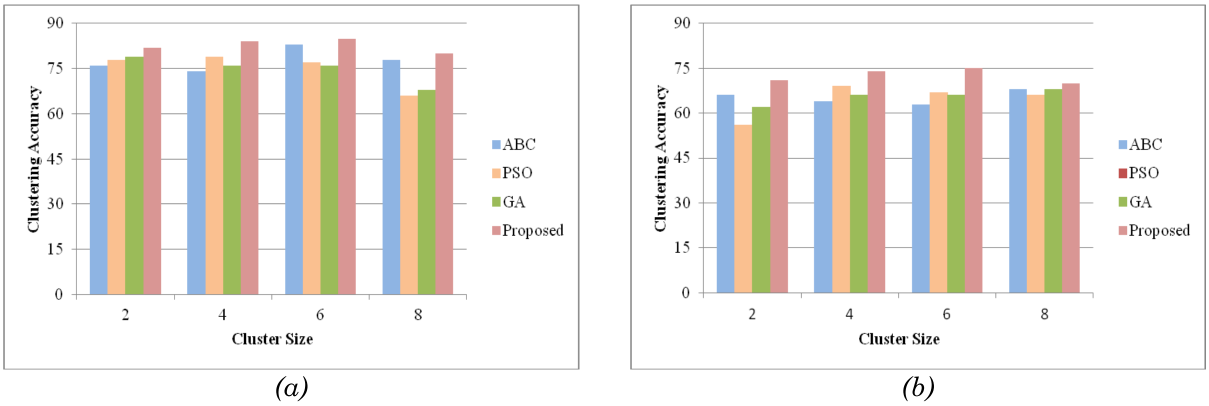


Fig. 7. *Performance evaluation of Clustering Accuracy (a)20 Newsgroup (b)Reuter*

In Fig 7, the performance evaluation of clustering accuracy on the basis of the cluster size is exhibited. The clustering accuracy plays an significant role to confirm the performance of the proposed method. Hence, Fig 7 (a) shows the performance of text clustering by 20 newsgroups. Fig 7 (b) depicts the maximum accuracy value offers the enhanced text clustering performance by the proposed algorithm and Wordnet ontology.

6. Conclusion

In this paper, the text clustering algorithm exploiting proposed Enhanced CSO algorithm and semantic word processing. The main aim of this algorithm was to produce the clusters from the input document. Initially, the input documents undertake for the preprocessing step in that the stop terms were evaded. In addition, the stemmer algorithm was exploited to extract valuable keywords for the following steps. Subsequently, the wordnet ontology was exploited to the ensuing document to decide the hyponyms and synonyms for each keyword. For every keyword the frequency was estimated for extracting the text feature by the clustering approach. Because of the dimension issue of feature space, the entropy measure to choose the significant features was exploited. The new Enhanced CSO algorithm was presented once the features were extracted. Hence, the proposed algorithm was exploited to discover the centroid to cluster the text documents. The experimentation outcomes were validated and performance was analyzed by Jaccard, clustering accuracy, and rand coefficient.

References

[1] R. Janani, S. Vijayarani,"Text document clustering using Spectral Clustering algorithm with Particle Swarm Optimization",Expert Systems with Applications, Volume 134, 15 November 2019, Pages 192-200.
[2] S. Yang, G. Huang and B. Cai, "Discovering Topic Representative Terms for Short Text Clustering," IEEE Access, vol. 7, pp. 92037-92047, 2019.
[3] Mohamed Atef Mosa,"Real-time data text mining based on Gravitational Search Algorithm Expert Systems with Applications", Volume 137, 15 December 2019.
[4] Elizaveta K. Mikhina, Vsevolod I. Trifalenkov,"Pages 117-129Text clustering as graph community detection",Procedia Computer Science, Volume 123, 2018, Pages 271-277.
[5] Caiyan Jia, Matthew B. Carson, Xiaoyang Wang, Jian Yu,"Concept decompositions for short text clustering by identifying word communities", Pattern Recognition, Volume 76, April 2018, Pages 691-703.
[6] Mosa, M. A. (2017). How can ants extract the essence contents satellite of social net- works . LAP Lambert Academic Publishing ISBN: 978-3-330-32645-3 . Mosa, M. A. (2018). Swarm intelligence optimization techniques based multiple types of text summarization with its satellite contents. Knowledge-Based Sys- tems .
[7] Mosa, M. A. , Hamouda, A. , & Marei, M. (2017a). Ant colony heuristic for user-con- tributed comments summarization. Knowledge-Based Systems, 118 , 105–114 .
[8] Mosa, M. A. , Hamouda, A. , & Marei, M. (2017b). Graph coloring and ACO based sum- marization for social networks. Expert Systems with Applications, 74 , 115–126 .
[9] Becker, H. , Naaman, M. , & Gravano, L. (2011). Selecting quality twitter content for events. In ICWSM (p. 11) .
[10] Bollen, J. , Mao, H. , & Pepe, A. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In ICWSM: 11 (pp. 450–453) .
[11] Harabagiu, S. M. , & Hickl, A. (2011). Relevance modeling for microblog summariza- tion. ICWSM .
[12] Hu, M. , & Liu, B. (2004). Mining and summarizing customer reviews. In Proceed- ings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168–177). ACM .
[13] Wahiba Ben Abdessalem Karaa, "A New Stemmer to Improve Information Retrieval", International Journal of Network Security & Its Applications (IJNSA), vol. 5, no. 4, pp. 143-154, 2013.

[14] Y. Sharafi,M . A. Khanesar and M. Teshnehlab," Discrete binary cat swarm optimization algorithm," 2013 3rd International Conference on Computer, Control & Communication (IC4), Karachi, 2013, pp. 1---6.
[15] S. G. Wang, F. F. Zhou, and F.J. Wang, "Effect of inertia weight w on PSO-SA algorithm," International Journal of Online Engineering, vol.9, no. 6. pp. 87-91, 2013.
[16] Neenavath Veeraiah and Dr.B.T.Krishna,"Intrusion Detection Based on Piecewise Fuzzy C-Means Clustering and Fuzzy Naive Bayes Rule",Multimedia Research, Volume 1, Issue 1, October 2018.
[17] Renjith Thomas and MJS. Rangachar,"Hybrid Optimization based DBN for Face Recognition using Low-Resolution Images"Multimedia Research, Volume 1, Issue 1, October 2018.
[18] Lewis DD (2015) Reuters-21578 text categorization test collection, http://www.daviddlewis.com/resources/testcollections/reuters 21578/.
[19] The 20 newsgroup dataset for text clustering, http://www.csmining.org/index.php/id-20-newsgroups.html.