DOI: http://dx.doi.org/10.26483/ijarcs.v9i1.5359

ISSN No. 0976-5697

Volume 9, No. 1, January-February 2018

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

TEXT DOCUMENT CLUSTERING USING ARTIFICIAL BEE COLONY WITH BISECTING K-MEANS ALGORITHM

Mrs. R.Janani
Ph.D Research Scholar
Dept. of Computer Science and Engineering
Bharathiar University
Coimbatore,India

Dr. S.Vijayarani
Assistant Professor
Dept. of Computer Science and Engineering
Bharathiar University
Coimbatore,India

Abstract: Recently, document clustering with optimization techniques has gained the attention of many researchers, especially those who are dealing with a huge volume of documents. The main goal of document clustering is to place the documents with similar content in one group, and the documents with dissimilar contents in another group. Document clustering with optimization algorithm achieves the global optimal solution. The main aim of this research work is to cluster the documents based on their content. In order to perform this task, this research work proposes a new hybrid algorithm called Artificial Bee Colony with Bisecting K-Means (ABC-BK). The proposed algorithm was verified with the benchmark dataset in contrast to the widely used document clustering algorithms. Experimental results show that the proposed algorithm gives a better performance compared to the standard ABC clustering algorithm, K-means and the Bisecting K-means algorithm.

Keywords: Document Clustering, ABC Algorithm, K-means, Bisecting K-means, ABC-BK

I. INTRODUCTION

Document clustering is most probably used to retrieve the collection of documents with similar contents in the area of information retrieval to enrich the document retrieval process. It is the procedure of dividing documents into one group, which is called clusters, in order that the documents in one group have the related contents [1]. In this research work, the most popular clustering algorithms are taken for experimentation.

Optimization algorithms may be both deterministic and stochastic in essentiality. Former techniques are used to solve optimization issues requires the massive computational resolves, which generally tend to flop because the problem size will gradually increase [2]. This is the inspiration for retaining bio inspired stochastic optimization algorithms as computationally efficient replacements to deterministic method [3]. Bio inspired algorithms are used to develop the global models to solve the problems. In this, the swarm algorithms are imagining the rush of feathered creatures and school of fishes. The Artificial Bee Colony (ABC) calculation imagining the searching conduct of honey bees.

This paper is organized as follows, section II explains the review of literature and section III presents the methodology of this research work. Experimental results are given in section IV and section V describes the conclusion of this research work.

II.LITERATURE REVIEW

In [4] the authors were analyzed the performances of the basic Artificial Bee Colony, Harmony Search, Bees algorithms and Improved Bees algorithm. These algorithms were compared and given the solutions on unimodal and multimodal benchmark problems. The performance of ABC algorithm is similar to the stated algorithm and ABC algorithm is effectively working to find the solution for the engineering problems with higher dimensionality.

In [5] they have presented the clustering center optimization and improving the accuracy. In this the initialclustering focuses of customary bisecting K-means algorithm are haphazardly chosen. In their work, they proposed an enhanced bisecting K-means algorithm which depends on the consequently conclusive K esteem and the improvement of the group focus. The examination comes about on UCI database demonstrated that the calculation can successfully avoid the commotion focuses and anomalies, and enhanced the exactness of clustering comes about.

In [6] they have displayed a novel transformative algorithm, called artificial bee colony (ABC), to enhance the capacity of k-means for finding global ideal clusters in nonlinear partitioned cluster issues. The proposed technique is the stage of k-means and ABC algorithms, called kABC, found the better group partitions. The reenactment comes about demonstrating that the mix of ABC and k-means method has greater capacity to scan for global ideal arrangements and greater capacity for passing neighborhood.

III. METHODOLOGY

The main objective of this research work is to retrieve the documents which have the higher similarity. In order to perform this clustering task, this research work proposes a new hybrid algorithm called Artificial Bee Colony with Bisecting K-Means (ABC-BK). Figure 1 shows the architecture of this research work.

A. Document Corpus

The collection of text documents is denoted as a document corpus. In this research work, the existing and proposed algorithms are verified with the following data set. They are Reuter's dataset, 20 newsgroup dataset and BBC dataset. These document datasets have been mostly used for evaluating the performance of document clustering and classification. These datasets are collected from different sources which contains newspaper articles, newsgroup posts and the remaining being academic news from the BBC news channel. The comprehensive summary of the data set used for experimentation is given in Table I.

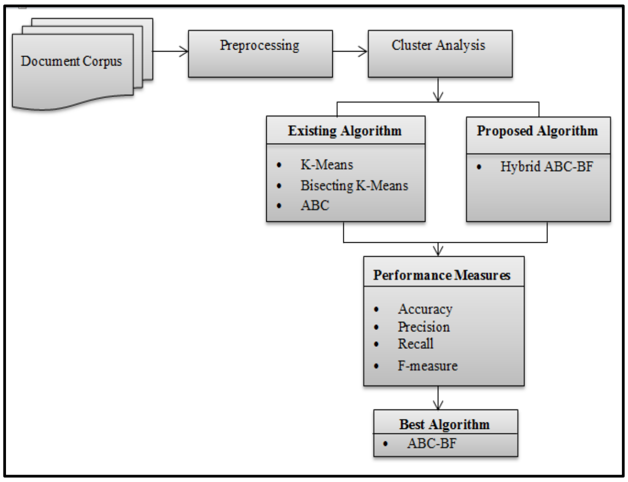

Figure 1: Methodology

Table I: Summary of Dataset

Dataset	*Source*	*Number of Documents*	*Number of classes*	*Number of Words*
re0	Reuters	1504	13	11465
20news	20news group	18828	20	28553
BBC	BBC News	2225	5	39558

B. Preprocessing

Document preprocessing is an important step in the process of document classification, clustering, topic identification, etc., Ininformationretrievalthepreprocessing techniques are applied to the precise dataset to extract the significant information from unstructured documents and to moderate the size of the dataset [7]. This process will increase the proficiency of the document clustering system. In this research work, stemming, stop word removal, numbers and punctuation removal techniques are used to extract the significant knowledge.

C. Clustering Algorithms

Clustering algorithms are commonly used in the field of text mining and machine learning. In text mining, the clustering algorithms are used to group the documents based on their content similarity of documents [1]. In this research work, the most popular clustering algorithms are taken for experimentation. They are k-means, BisectingK-means and ABC clustering algorithms.

a. K-Means

The k-means clustering algorithm is well known and competent partition algorithm which is used for large document collections. The main objective of this algorithm is to categorize the documents in a predefined number of clusters based on the document attributed or its features. This algorithm has been improved with other optimization techniques recently to improve the clustering accuracy [2]. Assume a fixed number of clusters; k-means are used to discover the partition of the documents established in a set of frequent features specified by the distance metrics. These metrics are used to define how the clusters are determined.

Algorithm: K-means

```
Input: A documents of n elements D = {$d_1$, $d_2$, …,$d_n$} and k the number of predefined clusters, $V_i$ the model vector
Output: k clusters of document
Repeat
Step 1: Assign each $d_n$to the nearest model vector $V_i$, cluster k contains documents $d_n$which are closest to $V_i$.
Step 2: Amend the new centroids rendering to,
        $V_i=\frac{1}{|k|}\sum_{\forall d_n \in K} d_n$
Unless convergence is achieved.
```

b. Bisecting K-Means

The bisecting k-means algorithm is an enhanced form of the k-means clustering algorithm. The essential thought of bisecting k-means is to achieve the quantity of C clusters, partitioned the arrangement of all focuses into two groups c1, c2 and choose one of these clusters (c1or c2) to part and so on, preceding C groups has been made [8].It is the powerful method to diminish the dimensionality of featured vector for classification and clustering. To choose which cluster to split, there are various ways are available [9]. In this research work, the criterion established on both size and the lowest sum of squared error method used.

Algorithm: Bisecting K-means

```
Input: C - Number of Clusters, D – Number of Documents.
Output: C Clusters of documents
Step 1: Initialize the list of clusters ($L_s$(c)) to contain the
        cluster consisting of all points and n=1 the number
        of iterations
Step 2: Repeat
Step 3: Choose the appropriate cluster from the $L_s$(c)
Step 4: For n= 1 to Cdo
Step 5: Bisect the number of clusters using basic k-means Bisect(c)
Step 6: End for
Step 7: Select the two clusters from $L_s$(c) with the lowest
        total sum of the squared error (SSE)
Step 8: add Bisect(c) ->$L_s$(c)
Step 9: until the $L_s$(c) containC clusters
```

c. Artificial Bee Colony (ABC)

The swarm intelligence algorithm used to be proposed by Karaboga among the year 2005. This algorithm is stimulated by means of scavenging behavior over the bee colonies [10]. To consign the solution regarding optimization problems, at that place is quite a few strategies are integrated along the algorithm itself. The solution about the fitness function is signified namely the fluid quantity regarding a food sources [11]. In accordance with this ABC algorithm, it consists of at that place are three kinds of bees such as, employed bees, onlooker bees and scout bees. Employed bees realize the unique food sources those have travelled earlier than and provide the virtue information as regards the food sources to the onlooker bees. Onlooker bees are adoption the information about the food sources and according to decide a food source after exploit based over the facts given with the aid of the employed bees. Scouts bees are ancient after search randomly between the environment in rule in accordance with find out a new food source. In that algorithm, half of the colony encompasses engaged bees and other half includes the onlooker bees [12,13].

Algorithm: Artificial Bee Colony

```
Input: $L_n$ , $U_n$ is the lower and upper bounds of n^th parameter
        K is the randomly selected food sources
        $\varphi$is a random number within the range [-1,1]
        N is the new food source on dimension d
Step 1: Initialize the food sources
    $S_{x,n} = L_n + Rand(0,1)(U_n - L_n)$
Step 2: Evaluate the fitness of food sources
    $N_{x,n} = d_{x,n} + \varphi\ (d_{x,n} - d_{y,n})$
    // Employed Bees
Step 3: Each bee in this phase yields the new food sources.
Step 4: Calculate the fitness function using Step 2
Step 5: For selection process, greedy search method is
        applied
Step 6: Calculate the probability of the food sources.
    $P_x = \frac{fitness_x}{\sum_{j=1}^{N} fitness_n}$
    // Onlooker Bees
Step 7: Choose the food sources based on the probability
        value based on employed bees
Step 8: Produce new food sources
Step 9: Calculate the fitness function using Step 2
Step 10: For selection process, greedy search method is
        applied
    // Scout Bees
Step 11: Swap the new food sources.
Step 12: New source will be produced and the counter will
        be 0
Step 13: Select the best food source.
Step 14: These three bees will be repeated until the best
        source will be found.
```

d. ABC-BK

As stated in the above sections, the existing clustering algorithms can generate the local optimal solution [14]. Those algorithms are extremely depending on the initialization of centroids and it converges after the number of iterations. Hence, in this research work we propose a combinational algorithm (ABC-BK) which uses the facts of bisecting algorithm and ABC algorithm for giving the solution of clustering problems [15]. This proposed algorithm is an independent of cluster centroids and also it leads the global optimal solution. In this proposed algorithm, the food particles in the pursuit foundation implies the group centroids or it signifies the solution for document clustering. The position P_i of the food source are assembledas

$$P_i = c_{x1}, c_{x2}, \dots\dots, c_{xk} \qquad (1)$$

where k is the total number of clusters and c_{xy} is the y^th cluster centroid of x^th food source. The proposed algorithm can briefly explain as follows,

- Initialize the food source position randomly and use thebisecting k-means algorithm to find out the solution for document clusteringfor all the position of food sources.
- To calculate the fitness value of each group usingthe fitness formula
- Examinenew food sources and appraise the position of the food sources by using the employed bees phase.
- Apply the bisecting k-means algorithm and a greedy search to estimatenew fitness value. These values are to be compared with original fitness values. Best food sources will be circulated to onlooker bee's phase.
- Determine the likelihoodestimationsof food sources and amend their position. Again, the bisecting k-means algorithm and a greedy search algorithm can be smeared to perform the document clustering.
- Then figure it the new fitness values and also update it.
- Review the counter of food sources and creates a new food source in the particular search space.

Algorithm: ABC-BK

```
Step 1: Initialize the food sources; send employed bees to
        current food sources
    Counter = 0;
Step 2: Calculate the fitness function
        // Employed Bees Phase
Step 3: Apply the bisecting k-means and greedy search
        algorithm.
Step 4: Compute the new fitness function.
Step 5: Estimate the probability value of food sources
        // Onlooker Bee's Phase
Step 6: Select the food sources based on the probability
        value
Step 7: Again, apply the bisecting k-means and greedy
        search algorithm.
Step 8: Compute the new fitness function.
        // Scout Bee's Phase
Step 9: Check the limit of the parameter.
Step 10: Produce the new food sources
Step 11: Counter = counter +1
Step 12: repeat the steps 3 to 11 until to get the optimal
        solution.
```

IV.Results and Discussion

In order to perform this clustering task, there are four performance measures are used in this research work. They are accuracy, precision, recall and f-measure [16, 17]. The experimental result establishes that the proposed algorithm gives the improved clustering results than the existing algorithms.

- **Accuracy**: Higher value of cluster accuracy indicates better cluster predictive discrimination.
- **Precision**: It is considered as the fraction of pairs properly situate in the same cluster.

Precision =——— (2)

Where R_1 is relevant documents and R_2 is retrieved documents.

- **Recall**: It is computed as the part of actual pairs that were recognized.

Recall = ——— (3)

Where R_1 is relevant documents and R_2 is retrieved documents.

- **F-measure**: It is the harmonic mean of precision and recall. It calculated as following

F-measure = ———— (4)

Table II shows the recall values for all the four algorithms. From this, we inferred the proposed algorithm gives the better accuracy when compared to other existing algorithms.

Table II: Recall for Three Datasets

Dataset	*K-means*	*Bisecting k-means*	*ABC*	*ABC-BK*
re0	30.09	34.12	31.91	36.18
20news	24.14	24.57	25.14	26.07
BBC	27.36	29.78	29.12	34.15

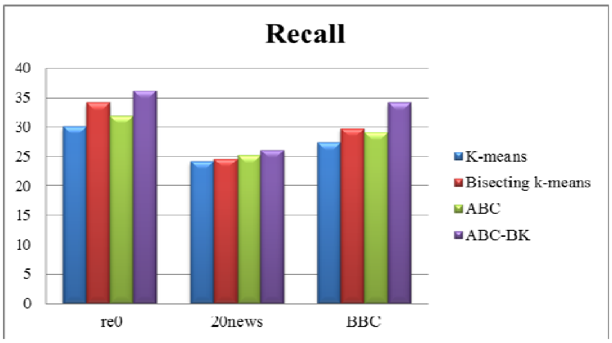

Figure 2: Recall

Table III shows the precision values for all the four algorithms. From this, we concluded the proposed algorithm gives the better accuracy when compared to other existing algorithms.

Table III: Precision for Three Datasets

Dataset	*K-means*	*Bisecting k-means*	*ABC*	*ABC-BK*
re0	47.15	49.78	49.10	54.18
20news	60.25	59.47	61.27	62.00
BBC	49.28	49.89	50.17	50.94

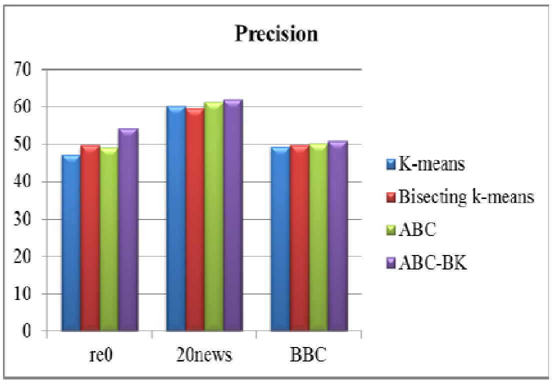

Figure 3: Precision

Table IV shows the f-measure values for all the four algorithms. From this, we concluded the proposed algorithm gives the better accuracy when compared to other existing algorithms.

Table IV: F-Measure for Three Datasets

Dataset	*K-means*	*Bisecting k-means*	*ABC*	*ABC-BK*
re0	36.73	40.48	38.68	43.38
20news	34.46	34.77	35.65	36.70
BBC	35.18	37.29	38.85	40.88

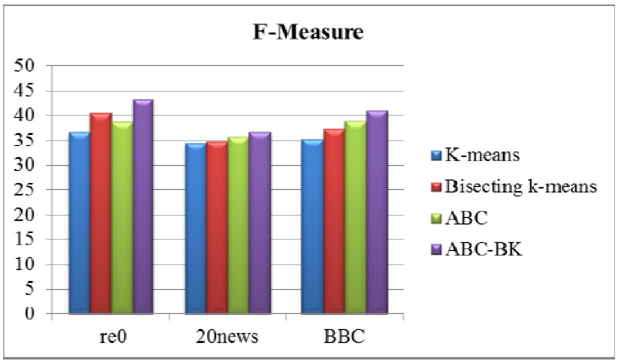

Figure 4: F-Measure

Table V shows the clustering accuracy for all the four algorithms. From this, we inferred the proposed algorithm gives the better accuracy when compared to other existing algorithms.

Table V: Accuracy for Three Datasets

Dataset	**K-means**	**Bisecting k-means**	**ABC**	***ABC-BK***
re0	69.21	70.18	71.25	80.97
20news	67.14	70.98	76.68	89.46
BBC	78.36	74.56	80.72	88.34

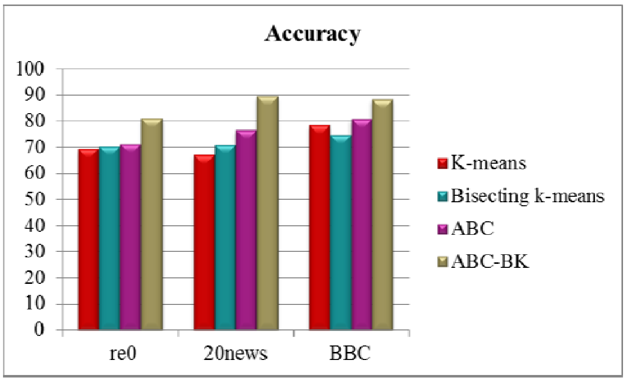

Figure 5: Accuracy

V. CONCLUSION

Document clustering plays an important role in unsupervised document organization, topic extraction, topic identification and information retrieval. This research work has developed a hybrid algorithm to attain the high clustering accuracy for handling unstructured documents. We explored the performances of k-means, bisecting k-means, ABC and ABC-BK algorithms. Based on the performance factors, the proposed algorithm converge the better optimal solutions. The algorithm has been implemented and verified on several real datasets. Also the ABC-BK algorithm takes the less control parameter to be altered with respect to the particular bees. In future, most effective algorithms to be developed to handle the huge volume of text documentsand to obtain the proper clustering accuracy.

VI. REFERENCES

[1] Nicholas O. Andrews and Edward A. Fox, “Recent developments in document clustering,” Technical report published by citeseer, pp. 1-25, Oct. 2007

[2]Jia, R., & Song, J. (2016). K-means optimal clustering number determination method based on clustering center optimization. Mcroelectronics and Computer, 5.

[3]Ji, X., Han, Z., & Li, K., et al. (2016). Application of the improved K means clustering algorithm based on density in the division of distribution network. Journal of Shandong University (Engineering Science Edition), 4, 41-46.

[4]D. Karaboga. An idea based on honey bee swarm for numerical optimization. Technical Report TR06, Erciyes University, Engineering Faculty, Computer Engineering Department, 2005.

[5] B. Basturk and D. Karaboga. An artificial bee colony (abc) algorithm for numeric function optimization. In IEEE Swarm Intelligence Symposium 2006, Indianapolis, Indiana, USA, May 2006.

[6]D. Karaboga and B. Basturk. A powerful and efficient algorithm for numerical function optimization: Artificial bee colony (abc) algorithm. Journal of Global Optimization, 39(3):459–471, 2007.

[7]Berry, M. (Ed.). "Survey of Text Mining: Clustering, Classification, and Retrieval". Springer, New York (2003)

[8] Liu, G., Huang, T., & Chen, H. (2015). Improved bisecting K-means clustering algorithm.Computer Applications and Software, 2.

[9]D. Karaboga and B. Basturk. On the performance of artificial bee colony (abc) algorithm. Applied Soft Computing, 8(1):687–697, 2008.

[10] K. R. Zalik, “An efficient k-means clustering algorithm,” Pattern Recognition Letters, vol. 29, pp. 1385–1391, July 2008.

[11] C. Zhang, D. Ouyang, and J. Ning, “An artificial bee colony approach for clustering,” Expert Systems with Applications, vol. 37, pp. 4761–4767, July 2010.

[12]W. Zou, Y. Zhu, H. Chen, and X. Sui, “A clustering approach using cooperative artificial bee colony algorithm,” Discrete Dynamics in Nature and Society, vol. 2010, pp. 16, October 2010.

[13]Pham, D.T. and Afify, A.A.: Clustering techniques and their applications in engineering. The Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science (2006)

[14] Nihal M. AbdelHamid, M.B. AbdelHalim, and M.W. Fakhr: Document clustering using Bees Algorithm. International Conference of Information Technology, IEEE, Indonesia (March 2013)

[15] Goldberg D.E.: Genetic Algorithms-in Search, Optimization and Machine Learning. Addison- Wesley Publishing Company Inc., London (1989)

[16]RekhaBaghel and Dr. RenuDhir, “A Frequent Concepts Based Document Clustering Algorithm,” International Journal of Computer Applications, vol. 4, No.5, pp. 0975 – 8887, Jul. 2010

[17] A. Huang, “Similarity measures for text document clustering,” In Proc. of the Sixth New Zealand Computer Science Research Student Conference NZCSRSC, pp. 49—56, 2008.