

A Hybrid Approach for Web Search Results Clustering using k-Medoids and Artificial Bee Colony Algorithm

Ahmed M. Al-Amdi

ahmedalamdy30@gmail.com

Basheer M. Al-Maqaleh

basheer.almaqaleh@tu.edu.ye

Faculty of Computers and Informatics,
Thamar University, Yemen

Faculty of Computers and Informatics,
Thamar University, Yemen

Abstract

Web Search Results Clustering (WSRC) is a useful method for organizing retrieved results into related groups. Existing WSRC approaches fail to cluster web search results effectively due to a number of difficulties such as over dependence on the initial cluster centers, easily falling into local optimum, and lack balancing between the exploration and exploitation capabilities due to the randomness and the limitations of their search strategies. In this paper, a hybrid approach for WSRC using *k-medoids* algorithm and Improved Artificial Bee Colony (IABC) algorithm is proposed to overcome these problems and to help users to identify effectively relevant information from a huge list of retrieved results. More precisely, the IABC algorithm is employed as the global search optimizer whereas *k-medoids* algorithm is employed as the local search optimizer. As a result, the proposed approach provide solutions that are locally and global balanced, thus the quality of the generated clusters is improved. The performance of the proposed approach is validated and tested on three well known datasets such as ODP-239, MORESQUE, and 20 Newsgroups. Three common evaluation measures are utilized to evaluate the performance of proposed approach, namely, precision, recall and F-measure. The obtained results of the proposed approach are satisfactory and achieved almost all the best results for all datasets in comparison with other competitive approaches.

Keywords: *Web Search Results Clustering, k-medoids Algorithm, IABC Algorithm, Snippets, Optimization.*

DOI:

<https://doi.org/10.59421/61btvb81>



THIS WORK IS LICENSED UNDER A
CREATIVE COMMONS ATTRIBUTION 4.0
INTERNATIONAL LICENSE.

I. Introduction

With the huge growth of information on the Internet, it has become very difficult for users to find relevant web pages [1]. The user starts looking for his required information from the top of the list and follows it down examining one result at a time, until the required information has been found [2]. Clustering is an essential data analysis technique that is used to organize a list of objects into coherent groups named clusters [3]. Text clustering is one of the methods that could be useful in achieving this goal by grouping Web Search Results (WSR) into topically coherent clusters [4].

The objective of the Web Search Result Clustering (WSRC) is to provide an easy way to the users to locate one or more related clusters that contains their requested information [5]. As a result, clustering of web documents has emerged as an interesting field of research within the scientific and academic communities working on Information Retrieval (IR) [1]. The common partitioning document clustering technique is *k-medoids* algorithm [6] which uses the document closest to the cluster center.

Artificial Bee Colony (ABC) was developed through simulation of intelligent foraging behavior of honey bees and it considers as one of the most common meta-heuristic optimization algorithms [7]. Basically, in ABC algorithm bees going to a food source already visited by them are employed bees while the bees looking for a food source are unemployed. Scout bees carry out search for new food sources and onlooker bees wait for the information from employed bees for food sources. There is one employed bee for every food source so, the exploitation process is performed by employed and onlooker bees phases whereas scout bees perform exploration of search space [8].

The search strategy of the original ABC algorithm has a good exploration capability and poor exploitation capability due to the randomness of its local search strategies [8,9]. The random method used to find a new food source in the original ABC algorithm causes uncertainty whether the search direction is a good or bad direction, therefore this situation leads to instability in the search strategy of the original ABC algorithm [10,11]. In the entire iteration process of the original ABC algorithm, the blindness and randomness make the algorithm more complex and sacrifices the accuracy of clustering results.

In this paper, a hybrid approach for WSRC using *k-medoids* and an Improved ABC (IABC) algorithm is proposed. In the proposed hybrid approach, the IABC algorithm is improved by incorporating effective search strategies in the original ABC algorithm phases to enhance its exploitation capability. As a result, the IABC algorithm is employed as the global search optimizer whereas *k-medoids* algorithm is employed as the local search optimizer. This is done by optimizing the randomly food source positions in the initialization phase of IABC algorithm in order to generate promising food source positions. In addition, the IABC algorithm is used to refine the obtained clusters from the *k-medoids* algorithm.

The rest of paper is organized as follows: Section II presents the related works, the proposed approach description is explained in Section III, computational results are presented in Section IV, and conclusion with future work are introduced in Section V.

II. Related Works

Several researchers have been proposed clustering approaches that can address the problems of traditional clustering approaches and can cope with WSRC challenges [4,5,10]. The k -means is a common clustering algorithm used in WSRC [11]. It uses the number of clusters k and initial clustering centroids to identify the related documents in each group or cluster using the similarity measurement. In [12], WSRC through density analysis based k -medoids algorithm was proposed. In this method, different attributes of the search results are cluster through on density analysis based k -medoids algorithm without defining categories in advance. The k -medoids clustering algorithm can efficiently improve the performance of the clustering techniques by selecting the document closest to the cluster center. In [13], a novel approach to text clustering using shift k -medoids was suggested. In this approach, the k -medoids and mean shift clustering algorithms are used to convert the large data into different cluster forms to extract and get the summarized data. A comparison of k -means and k -medoids clustering algorithms on Twitter data was presented in [14]. The results of this study indicate that the optimal number of clusters resulting from k -means and k -medoids are different.

In [10], WSRC using meta-heuristic algorithm was described. This algorithm is based on the usage of k -means to form the clusters of documents and the Balanced Bayesian algorithm to calculate the probability to decide in which cluster the document should be placed. A novel Particle Swarm Optimization(PSO) method for web documents retrieval was introduced in [15]. This method has been testified to possess superior performance in terms of response time, stability, robustness and it is well suited to handle both small and huge collections. In [16], the improved ABC algorithm, called AB2C, for clustering was introduced to converge faster that used the best solution obtained by bees to update their population and k -means algorithm is used to initialize the population. The proposed AB2C is applied to some UCI datasets and it obtains promising results.

In [17], a hybrid approach for WSRC based on a modified GA with k -means called G-K-M was proposed. This approach has provided the evaluation process of the enhanced and optimized web data clustering models for WSRC. A hybrid approach for web document clustering using k -means and ABC algorithm (WDC-KABC) was introduced in [18]. In this approach, there is no special improvement for the original ABC algorithm except its integration with k -means algorithm. In [19], a hybrid algorithm based on modified ABC and k -means algorithm (MABCKM) approach for data clustering was proposed. The experimental results prove the superiority of the MABCKM algorithm for data clustering applications. In [20], an enhanced clustering models with wiki-based k -nearest neighbors-based representation (wiki-KNN) for WSRC were proposed. These models have developed a new unsupervised distributed word representation scheme, where each word is represented by a vector of its semantically related words such as scheme expands snippets and user queries. The best results obtained when this representation used with k -medoids algorithm.

Different from the algorithms and approaches mentioned in this survey, the proposed approach is suggested as an attempt to incorporate the k -medoids algorithm with IABC algorithm in order to use the main merits of both algorithms for WSRC. More precisely, the proposed approach is improved in two directions; the first is the hybridization, and the second is the modification of search strategies in all phases of the proposed approach. In the proposed approach, the k -medoids algorithm is improved to overcome its problem such as initialization issues, over dependence on the initial clustering center, and easily falling into local optimum. Also, due to the blindness,

randomness and the limitations of search strategies in the original ABC algorithm some improvements have been suggested. Each phase of the proposed approach in the clustering stage is appropriately improved in order to generate the best solutions as will be explained in the next section.

III. The Proposed Approach Description

The proposed approach consists of four stages as following:-,

1. Pre-Processing Stage

All web documents are pre-processed to avoid wrong clustering and to remove noisy. In this stage, standard traditional Natural Language Processing (NLP) techniques [21] like tokenization, normalization, stop word removal and stemming are applied on a retrieved search result.

2. Document Representation Stage

In this stage, the bag-of-n-gram representation technique is used for data representation [22]. In n-gram with Term Frequency-Inverse Document Frequency (TF-IDF) representation matrix, every row represents a search result document, every column signifies an n-gram, and a cell contains the TF-IDF weighted value of an n-gram for a search result document. The TF-IDF value of a weight W_i is calculated as the product of two values, which correspond to the TF and IDF, respectively. It can be computed as under [22]:

$$W_i = TF_i \times IDF_i = TF_i \times \log\left(\frac{N}{n}\right) \quad (1)$$

where TF_i is the word frequency in document (Doc_i), IDF is the number of documents contains words, n is the number of documents containing the word and N the total number of documents. Each search results r_{ij} is represented as $r_i = (a_{i,0}, a_{i,1}, \dots, a_{i,m})$, where $a_{i,j}$ is the $TF - IDF_i$ in the search results r_i .

3. Clustering Stage

Clustering stage is the core stage and it has the following phases:

A. Generating initial food sources positions phase

The proposed approach starts with the random selection of food sources positions that represent possible solutions of a problem. The initial position of food sources is generated in the range of the boundary limit as below[7].

$$X_{i,j} = X_j^{Min} + rand(0,1)(X_j^{Max} - X_j^{Min}) \quad (2)$$

where $i = 1, 2, \dots, sn$ and $j = 1, 2, \dots, D$, sn represents the number of the food sources, D denotes the number of dimension of the problem, X_j^{Max} and X_j^{Min} denote the lower and upper limits for $X_{i,j}$ and $rand$ is random number in the range of 0 to 1.

The values generated using Equation (2) are considered as centroids, and used as input to k -medoids algorithm. More precisely, the proposed approach detects dynamically the initial value of k for creating initial centroids. This eliminates the problem of external specification of the

value k , which may lead to unwanted result if wrongly specified. The centroids created in this way are more specific and meaningful in the context of WSRC. The suggested initialization process not only reduces the randomness of colony initialization, but also reduces the sensitivity of k -medoids clustering to initial points.

B. Applying k -medoids algorithm for food sources optimization phase

Each food source in the search environment represents a set of centroids, which they used as input to k -medoids algorithm in order to cluster all produced positions. The k -medoids algorithm searches the local optimum solution in all possible initial solutions. In this case, the Euclidean Distance (ED) is used as a similarity measure to evaluate the best positions of food sources, which can be computed as under:

$$ED_{i,j}(X, Y) = \sqrt{\sum_{a=1}^n (X_{i,a} - Y_{j,a})^2} \quad (3)$$

where $X_{i,a}$ denotes the i,a -th data object or document, and $Y_{j,a}$ represents the j,a -th cluster center, $a = 1, 2, \dots, n$, and n is a set of data object.

After partitioning the data into different clusters, another similarity measure called intra-cluster distance ($Dist$) is applied to determine the goodness of the partitions by determining the strong association between the food sources "data objects or documents". It can be computed to measure the distance between two points P and Q as under [16]:-

$$Dist_{i,j}(P, Q) = \sum_{ia=1 \in X_{j,a}}^j (X_{i,a} - Y_{j,a}^2) \quad (4)$$

where $j = 1, 2, \dots, n$, $X_{i,a}$ represents the data object belonging to the j,a -th cluster centers, and $Y_{j,a}$ represents the j,a -th cluster centres.

The total intra-cluster distance is calculated by summing up the distance between each data object or document, and the cluster centre to which the data object belongs. So, smaller value of this measure denotes higher quality of clustering approach. An initial guess at the centres of the clusters is called P_{ij} and is computed as under [16]:

$$P_{i,j} = \frac{Dist_{i,j}}{\sum_{i=1}^n Dist_{i,j}} \quad (5)$$

The value of $\sum_{i=1}^n P_{i,j}$ is computed at each objects or documents and sort them in ascending order. Select k objects having the minimum value as initial group medoids, assign each object to the nearest medoid. Calculate the current optimal value as the sum of distance from all objects to their medoids. Find new medoids and replace the current medoid in each cluster by the object, which minimizes the total distance to other objects in its cluster. Assign each object to the nearest new medoid, and calculate new optimal value. If the optimal value is equal to the previous one, then stop the algorithm. Otherwise, repeat that. Here k is number of cluster centres positions that

would be obtained. These positions represent the possible locations of the food sources for the employed bee phase of IABC algorithm.

C. Employed bee phase

The output of *k-medoids* algorithm in the previous phase is used as input to the employed bee phase in the IABC algorithm. The employed bees phase deals with the search of new high quality food sources in the nearby locality of old food source using the following Equation (6) [7].

$$V_{i,j} = X_{i,j} + \phi_{i,j}(X_{i,j} - X_{k,j}) \quad (6)$$

where $V_{i,j}$ represents the position of the food source of documents set, $X_{i,j}$ gives the information about the neighborhood of the food source of documents, $\phi_{i,j}$ denotes a random variable in between $[-1, 1]$ to control the search region of the neighborhoods, and $X_{k,j}$ represents another solution randomly selected from the population that contain set of objects or documents and $k \neq i$. The fitness function (Fit_i) for food source in the population is computed as follows [9]:

$$Fit_i = \frac{1}{1 + F_i} \quad (7)$$

where Fit_i is the fitness of the *i*-th food source and F_i denotes a cost function for each food source of the population and it can be computed as under [16].

$$F_i = \frac{1}{n_j} \sum_{i=1}^{n_j} ED(X_j, C_i) \quad (8)$$

where n_j denotes the number of data instances that are used to normalize the sum, C_i is the class to which data instances belong and X_j is the generic instance. A greedy selection is performed between the $V_{i,j}$ and the $X_{i,j}$, and the better one is selected according to its fitness value. The probability value for each food source is computed as under [16].

$$P_i = \frac{Fit_i}{\sum_{j=1}^{sn} Fit_j} \quad (9)$$

where Fit_i is the fitness of the *i*-th food source "document" and j is the total number of clusters centers. An onlooker bee selects a food source depending on the probability value P_i according to the quality of the nectar amount.

D. Onlooker bee phase

Due to the blindness and randomness of the local search strategies in original ABC algorithm, which cause poor exploitation capability. The proposed approach suggests a new local search mechanism based on the social behavior of PSO that enables the onlooker bee to exploit the nectar amount to generate the more promising food source position. This is done by using the following Equation [15]:-

$$V_{i,j} = X_{i,j} + \phi_{i,j}(X_{i,j} - X_{k,j}) + \varphi_{i,j}(Y_j - X_{k,j}) \quad (10)$$

where Y_j represents j -th instance of the global best solution and $\phi_{i,j}$ is a uniform random number in $[-1,1]$.

By applying this new mechanism for local search, the IABC's blind search in the onlooker bee phase has turned into a smarter search in the search space. Along this, the parameter $\varphi_{i,j}$ also plays a significant role in Equation (10) for maintaining the balance between exploration and exploitation of the candidate solution search. The fitness function for each food source is computed again using Equation (7). The fitness of the solution is represented as the nectar amount of a food source. The more nectar the food source contains, the larger probability the onlooker bee will choose it.

E. Scout bee phase

In case the position of food source is not improved up to a user defined parameter called limit, the food source is abandoned and the corresponding employed bee becomes a looker (scout) bee, whose responsibility is to explore randomly search the whole environment. The limit parameter controls the employed bees whose food should be abandoned. In original ABC algorithm, a new position of food source is randomly generated in the whole search space using Equation (2). Whereas the proposed approach enhances the exploitation capability by generating a better position in scout bee phase instead of random position using a modified Hooke and Jeeves search method as follows [19]:-

$$X_{new} = X_{best} + rand[0,1](X_{best} - X_{current}) \quad (11)$$

where $X_{current}$ is the current position of candidate solution, and X_{best} is the best ever position achieved by this candidate solution. These phases would be executed until the termination criteria are satisfied. As a result, the best food source solutions are obtained as an output.

4. Evaluation Stage

The performance of the proposed approach is evaluated by using the standard IR measures such as precision, recall and F-measure. These measures are defined as under:-

- Precision aims to evaluate the cluster based on the number of correct retrieved documents out of the total number of retrieved documents in all clusters. It is computed as under.

$$Precision = \frac{(Number\ of\ correct\ retrieved\ documents)}{(Total\ number\ of\ retrieved\ documents)} \quad (12)$$

- Recall aims to evaluate the cluster based on the number of correct retrieved documents out of the total number of correct documents. It is computed as follows:-

$$Recall = \frac{(Number\ of\ correct\ retrieved\ documents)}{(Total\ number\ of\ correct\ documents)} \quad (13)$$

- F-measure is the complete clustering result of the documents to obtain the accuracy of the

clustering results. It is based on precision and recall, and it is computed as under:-

$$F - measure = \frac{(2 \times Precision \times Recall)}{(Precision + Recall)} \quad (14)$$

Obviously, the bigger values of these measures suggest a good quality of clustering. The pseudocode of the proposed approach is given below in Figure 1.

SRL: dataset contains n search result snippets or documents.

1. Document Pre-processing Stage.
 2. Document Representation Stage.
 3. Clustering Stage.
-

A. IABC Initialization and k -medoids algorithm for food sources optimization phases:

Initialize IABC algorithm parameters such as *food sources()*, *limit*, *maximum cycle number*, *colony size* and *upper & lower boundary*;

For $i = 1$ to k /* k is the total number of food sources */

Randomly assign the population of food sources (clusters medoids) using Equation (2) to k -medoids algorithm;

Apply k -medoids algorithm to evaluate the better positions of food sources using Equations (3), (4), and (5);

Send the optimized food sources to the employed bee phase;

End For

B. Employed bees phase:

Iteration = 0;

While (Stopping condition is not met)

For $i = 1$ to k /* i represents the i -th employed bee and k represents the total number of food sources */

Discover a new food source in the neighboring of old one using Equation (6);

Compute fitness of new generated food source using Equations (7) and (8);

Apply a greedy selection to choose the best food source using Equation (9); **End**

For

C. Onlooker bees phase :

For $i = 1$ to k /* i represents the i -th onlooker bee and k represents total number of food sources */

If (rand () < Pi) / Pi denotes probability associated with i-th food source */*

Send the onlooker bee to exploit the new food source using Equation (10);

Compute the fitness of the new food source using Equations (7) and (8);

Apply a greedy selection to choose the best food source using Equation (9);

Else i=i+1;

End If

End For

D. Scout bee phase:

If the food source is not improved further using limit;

Send a scout bee to generate a new food source position using Equation (11);

End If

Memorize the best solution achieved so far.

```
Iteration = iteration + 1;
```

End While

Output the final cluster medoids denote to clusters centers.

4. Evaluation Stage

Apply evaluation measures (precision, recall, and F-measure) using Equations (12), (13), and (14) respectively to evaluate the quality of generated clusters.

Figure 1: The pseudocode of the proposed approach.

IV. Computational Results

The performance of the proposed approach is tested and validated on some benchmark real world datasets that used in evaluating WSRC algorithms. Several experiments are carried out to assess the proposed approach performance for WSRC. In addition, its performance is evaluated and compared with other related approaches such as *k-means* and *k-medoids* [23], PSO [15], hybrid PSO with *k-means* [24], *Gen-G-K-M* [17], WDC-KABC [18], and wiki-KNN [20] based on common IR evaluation measures like precision, recall, and F-measure.

The values of control parameters of the IABC algorithm are to be set as follows: The colony size of IABC was set to 20, the maximum cycle number was set to 2000, the abandonment limit was set to 10, the value of $\emptyset$ was set to 0.75, the upper bounce was set to 5, and lower bounce was set to 0.01. All experiments are performed on a laptop with C# 2012 windows10, 4GB RAM, and 2.4 GHZ core i5 processor. The description of the selected datasets is demonstrated below.

A. Datasets Description

The proposed approach is tested and validated using the following datasets:-

1. Open Directory Project (ODP-239) Dataset

The ODP-239 dataset is generated from open directory project with 25580 retrieved snippets, which means about 100 web search results/snippets for each query of 239 queries [25]. Table 1 shows the statistical information of ODP239 dataset.

Table 1: Statistical information of ODP239 dataset.

Number of used queries	239
Number of clusters for all queries	2285
Average number of search snippets (retrieved for all queries)	25580
Average number of clusters per query	9.5607 (min 6, max 10)
Average number of retrieved search results per query	107.0293 (min 98, max 131)
Average number of retrieved documents per cluster	(min 4, max 94)

2. More Sense-Tagged Queries (MORESQUE) Dataset

The MORESQUE dataset has developed for evaluating IR results clustering [26]. This dataset consists of 114 queries or topics, each of which has a set of subtopics and a list of 100 top-ranking web documents. It includes queries with different lengths of 2, 3, and 4 as well as the top 100 Yahoo! retrieved results for each query totaling 11,261 snippets. Table 2 shows some statistical information of MORESQUE dataset.

Table 2: Statistical information of MORESQUE dataset.

Number of queries	114
Number of subtopics clusters	753
Number of retrieved snippets	11261
Average number of clusters per query	6.6
Average number of retrieved documents per query	100

3. 20 Newsgroups Dataset

The 20 Newsgroups dataset is a collection of approximately 18,828 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroup[27]. Table 3 shows some statistical information of 20 Newsgroups dataset.

Table 3: Statistical information of 20 Newsgroups dataset.

Number of unique documents	18828
Number of categories	20
Number of unique words after removing the stop words	71830

The experimental results from these three datasets are demonstrated below.

B. Experimental Results

The obtained results of the proposed approach in terms of precision, recall and F-measure for WSRC from three selected datasets as shown in Table 4.

Table 4: Evaluation results of the proposed approach over three datasets.

Dataset Name	Precision	Recall	F-measure (%)
ODP-239	90.10%	89.10%	89.60%
MORESQUE	90.10%	92.10%	91.09%
20 Newsgroups	91.18%	93.43%	92.29%

The precision rate ranges from 90.10% to 91.18% and these results assume that the documents clustered together are relevant out of the total count of clustered documents. Also, on analyzing the experimental results of recall rate with respect to different datasets. It has proved that the recall rate lies between 89.10% to 93.43%. Thus, the recall rate shows that the proposed approach clusters effectively the relevant documents together out of the total count of actual relevant documents. F-measure ranges from 89.60% and 92.29%. The values of this measure are directly proportional to the quality of the cluster solution. Thus, the quality of clusters is found to be satisfactory in the proposed approach as the high values of F-measure indicate how the clusters are spatially well separated.

Additionally, the proposed approach is almost capable of automatically finding the actual number of clusters presents in the WSR with high values of precision and recall. Table 5 shows the final number of generated clusters by the proposed approach and the defference to the actual number of clusters presents in the ODP-239 dataset.

Table 5: Sample of ODP 239 dataset queries, actual number of clusters and number of clusters generated by the proposed approach.

Query No..	No. of actual clusters	No.of generated clusters	Difference
1	10	8	2
2	10	9	1
3	10	10	0
4	8	8	0

5	9	9	0
6	7	6	1
236	10	9	1
237	10	10	0
238	10	9	1
239	10	10	0

C. Comparative Study

The performance of the proposed approach is compared with other related approaches in terms of precision, recall and F-measure as shown in Table 6.

Table 6: Comparison results of the proposed approach and other competitive approaches.

Approach	ODP239			MORESQUE			20 Newsgroups		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
<i>k-means</i>	73.80%	73.66%	73.73%	78.01%	81.32%	79.63%	78.60%	82.80%	80.65%
<i>k-medoids</i>	78.07%	78.73%	78.40%	79.91%	82.53%	81.20%	81.85%	82.90%	82.37%
PSO	80.20%	80.90%	80.55%	80.90%	81.95%	81.42%	81.88%	83.80%	82.83%
PSO+ <i>k-means</i>	81.75%	82.85%	82.30%	82.20%	83.50%	82.85%	84.50%	86.40%	85.45%
G-K-M	87.39%	88.59%	87.99%	87.73%	88.50%	88.11%	90.05%	90.61%	90.33%
WDC-KABC	83.50%	84.90%	84.19%	89.80%	90.16%	89.98%	91.80%	91.61%	91.70%
wiki-KNN	85.13%	84.35%	84.74%	87.66%	86.48%	87.07%	84.95%	86.70%	85.82%
The Proposed Approach	90.10%	89.10%	89.60%	90.10%	92.10%	91.09%	91.18%	93.43%	92.29%

The comparison results on these three datasets show that clustering performance varies from one dataset to other dataset. The best results based on the overall evaluation measures are the highest. A good retrieved system should have high precision for most levels of recall. The hybridization in the proposed approach creates a diverse cluster and selects the high-quality clusters from candidate clustering solutions, thereby leading to better clustering approach. On analyzing the comparative study results with respect to different datasets, it is proved that the proposed approach outperforms all other related approaches with higher rates of precision, recall and F-measure in all three datasets

Figures 2, 3, and 4 present the comparative performance based on precision, recall, and F-measure for ODP239, MORESQUE, and 20 Newsgroups datasets respectively.

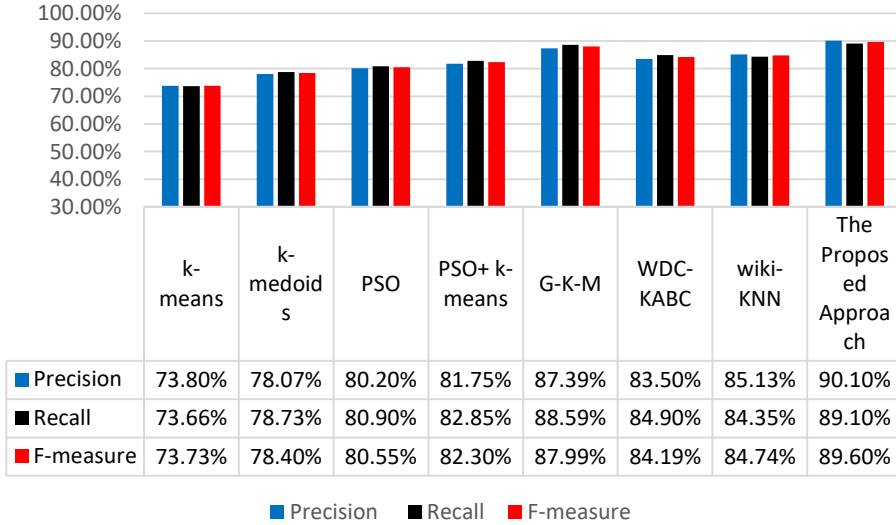


Figure 2: The comparative performance analysis based on the ODP239 dataset.

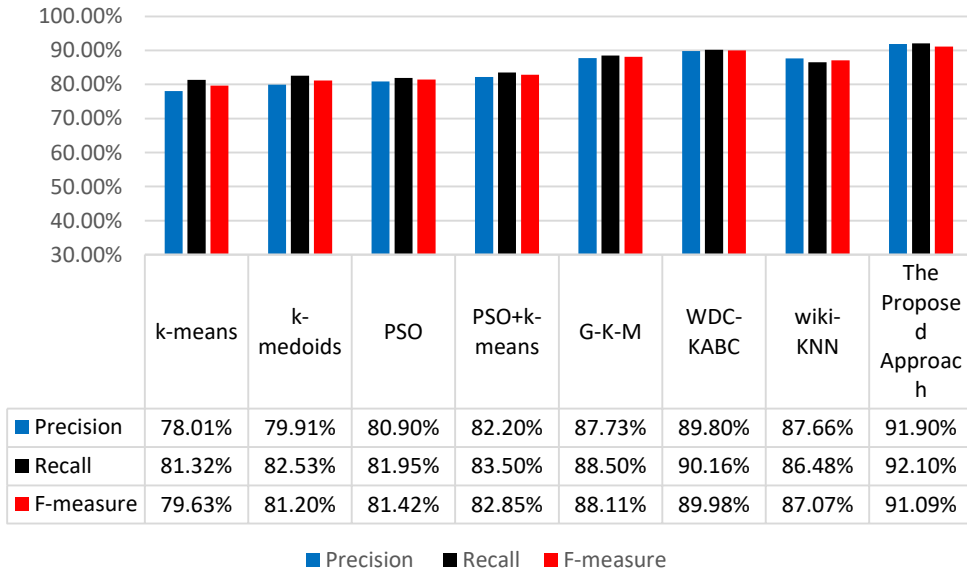


Figure 3: The comparative performance analysis based on the MORESQUE dataset.

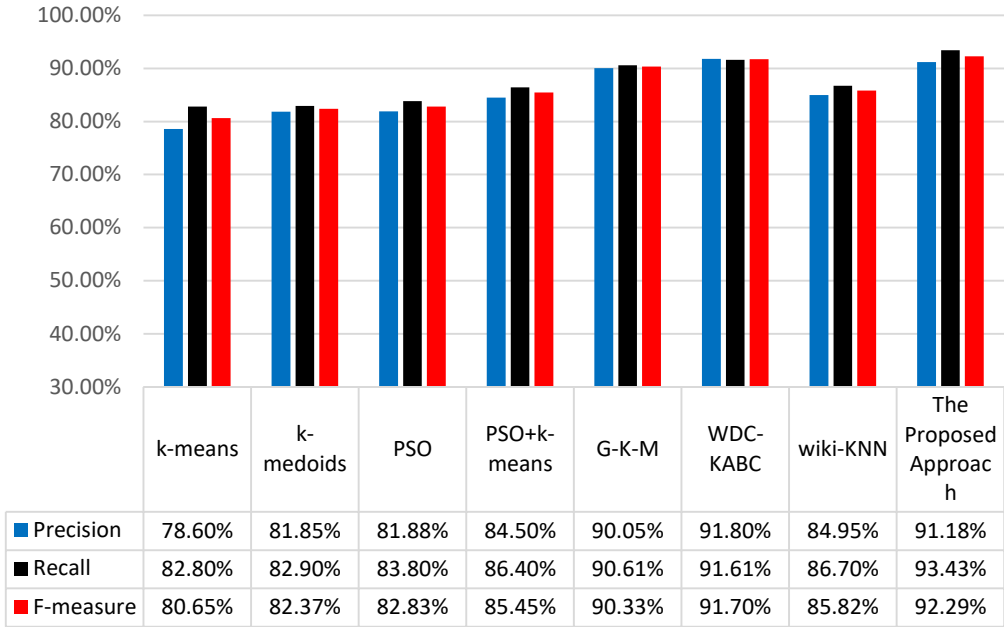


Figure 4: The comparative performance analysis based on the 20 Newsgroups dataset.

V. Conclusion and Future Work

In this paper, a hybrid approach which is a combination of *k-medoids* algorithm and IABC algorithm is proposed. The IABC algorithm is used as a global search optimizer and *k-medoids* algorithm is employed as the local solution optimizer. This hybridization resulted in strong exploration and exploitation, thereby effectively leading the search process and obtaining high quality solutions.

The performance of the proposed approach is validated using three benchmark real-world datasets and the obtained results were satisfactory, and showed the effectiveness of the proposed approach. The comparative study in all three datasets proved that the proposed approach outperforms all other related approaches with higher rates of precision, recall and F-measure. One of the most important future research directions would be developing the proposed approach to update the set of generated clusters when new dynamic snippets are inserted, deleted or modified by applying the incremental learning concept.

References

1. P. Verma, N. Kesswani, A. Jain, and P. K. Mishra, *Web Mining: Analyzing Traversal Patterns*, DeepMisti Publication, 2024.

2. P. S. Sharma, D. Yadav, and R. N. Thakur, "Web Page Ranking using Web Mining Techniques: A Comprehensive Survey", *Mobile Information Systems, Hindawi*, 2022.
3. M. A. Al-Mojahed and B. M. Al-Maqaleh, " DEADC: Density Extending Algorithm for Data Clustering", *2nd International Conference on Emerging Smart Technologies and Applications(eSmarTA)*, Ibb, Yemen, pp. 1-8, 2022.
4. M. H. Ahmed, S. Tiun, N. Omar, and N. S. Sani, " Short text clustering algorithms, applications and challenge: A survey", *Applied Sciences, MDPI*, Vol. 13, Issue 1, 2023.
5. A. J. Obiad, T. Chatterjee, and A. Bhattacharya, " Semantic Web and Web Page Clustering Algorithms: A Landscape View", *EAI Endoresd Transactions on Energy Web*, 8(33), pp. 1-14, 2020.
6. L. Kaufman, P. Rousseeuw, and Y. Dodge, "Clustering by Means of Medoids", *In Statistical Data Analysis Based on the L₁-Norm and Related Methods*, North-Holland, Amsterdam, pp. 405-416, 1987.
7. D. Karaboga, " An Idea Based on Honey Bee Swarm for Numerical Optimization", *Technical Report (TR06)*, Computer Engineering Department, Engineering Faculty, Erciyes University, 2005.
8. L. Abualigah, A. H. Gandomi, M. AbdElaziz, A. G. Hussien, A. M. Khasawneh, M. Alshinwan, and E. H. Houssein., "A Nature-Inspired Optimization Algorithms for Text Document Clustering—A Comprehensive Analysis ", *Algorithms*, 13(345), pp.1-32, 2020.
9. Y. Celik, " An Enhanced Artifical Bee Colony Algorithm Based on Fitness Weighted Search Strategy", *Automatika Journal*, 62(3), pp. 300-310, 2021.
10. R. Agarwal, S. Gawade, P. Pirale, and D. Dighe, " Clustering of Web Search Result using Meta-heuristic Algorithm", *International Journal of Engineering Research and Technology (IJERT)*, 7(3), pp. 318-319, 2018.
11. Q. Jin, N. Lin, and Y. Zhang, "K-Means Clustering Algorithm Based on Chaotic Adaptive Artificial Bee Colony", *Algorithms*, MDPI, 14(53), pp.1-23, 2021.
12. H. Hung, and W. Junzo, " Search Result Clustering through Density Analysis Based K-mediod Method", *International Conference on Advanced Applied Informatics(IIAI)*, pp. 155-160, 2014.
13. Mohit, A. C. Kumari, and M. Sharma, " A Novel Approach to Text Clustering using Shift k-Medoid", *International Journal of Scocial Computing and Cyber-Physical Systems*, 2(2), pp. 106-118, 2019 .
14. C. Oktarina, K. A. Notodiputro, and I. Indahwati, "Comparison of K-Means Clustering Method and K-Medoids on Twitter Data", *Indonesian Journal of Statistics and its Applications*, vol. 4, pp. 189-202, 2020.
15. C. Ramya, " A Novel PSO Methodolgy for Web Documents Retrival ", *International Journal of Computational Linguistics Resreach*, 10(3), pp. 67-75, 2019.

16. S. M., H. Hasheminejad, M. Vosoughian, and M. Zamini, " AB2C: Artificial Bee Colony for Culstering", *International Journal of Artificial Intelligence, CESER Publications*, 18(2), pp. 51-73, 2020.
17. B. Al-Attar, J. Allami, and T. A. Imeer, "A Hybrid Approach for Web Sreach Result Clustering Based on Genetic Algorithm with k -Means", *Journal of Theoretical and Applied Information Technology*, vol. 99, pp.2722-2733 , 2021.
18. M. G. Alam, and S. Baulkani, "A Hybrid Approach for Web Document Clustering using k -Means and Artificial Bee Colony Algorithm", *International Journal of Intelligent Engineering and Systems*, vol. 9, pp. 11-20, 2016.
19. A. Kumar, D. Kumar, and S. Jarial, "A Hybrid Clustering Method Based on Improved Artificial Bee Colony and Fuzzy C-Means Algorithm", *International Journal of Artificial Intellegence*, vol. 15, pp. 24-44, 2017.
20. A. S. Abdulameer, S. Tiun, N. S. Sani, M. Ayob and A. Y. Taha, "Enhanced Clustering Models with wiki-Based K-Nearest Neighbors-Based Representation for Web Search Result Clustering", *Journal of King Saud University-Computer and Information Sciences*, vol. 34, pp. 840-850, 2022.
21. R. Lee, *Natural Languge Processing: ATextbook with Python Implementation*, 2nd Edition, Springer, 2025.
22. Han. and M. Kamber, *Data Mining Concepts and Techniques*, 4th Edition , Morgan Kaufmann Pulpishers, 2022.
23. R. C. Balabantaray, C. Sarma, and M. Jha, "Document Clustering using k -Means and k -Medoids", *International Journal of Knowledge Based Computer System*, 1(1), 2013.
24. .S. Mehrotra, and A. Sharan, " Comparative Analysis of k -Means Algorithm and Particle Swarm Optimization for Search Result Clustering", *Smart Trends in Computing and Communications, Springer*, vol. 165, pp. 209-114, 2020.
25. C. Carpineto, and G. Romano, " ODP-239 Dataset, 2009, <http://credo.fub.it/odb239/>
26. R. Navigli, and G. Crisafulli, " Inducing Word Senses to Improve Web Search Result Clustering", *In Proceedings of the 2010 Confrence on Empirical Methods in Natural Language Processing (EMNLP 2010)*, MIT Stata Center, Massachusetts, USA, pp. 116-126, 2010.
27. L. Blake, and M. J. Merz, UCI Repository of Machine Learning Database [<http://www.ics.uci.edu/mllearn/MLRepository.html>], Irvine, CA: University of California, Department of Information and Computer Science.