WJCMS

Wasit Journal of Computer and Mathematics Science

Journal Homepage: https://wjcm.uowasit.edu.iq/index.php/WJCM

e-ISSN: 2788-5879 p-ISSN: 2788-5879

An Effective Algorithm to Improve the Accuracy of Recommender System based on Comments using Classification Techniques in Data Mining

Razieh Asgarnezhad[1], Ali Naseer Kadhim Alwali[1,2], Mhmood hamid sahar alsaedi[1] and Samer alwan zaboon albwhusseinsarr[1,2,*]

[1]Department of Computer Engineering Isfahan (Khorasan) Branch Islamic Azad University Is- fahan , Iran
[2]Ministry of Education Educational Directorate of Wasit province Wasit Iraq

*Corresponding Author: Samer alwan zaboon albwhusseinsarr

DOI: https://doi.org/10.31185/wjcm.Vol1.Iss1.27
Received: December 2021; Accepted: February 2022; Available online: March 2022

ABSTRACT: With the development of information systems, data has become one of the most important sources of organizations. Therefore, methods and techniques are needed to efficiently access data, share data, extract data from data, and use this information. By creating and expanding the Web and a significant increase in the volume of information and web development, the need for methods and techniques that can provide data efficiently and extract information from them is felt more than ever. Web mining is one of the areas of research that uses data mining techniques to automatically discover information from web services and documents. In fact, Web mining is a process of discovery of unknown and useful information from web data. Web mining methods are categorized into three types of web content exploration, exploration of Web structures, and exploration of the use of the Web, based on what type of data they are exploring. This research investigates the relationship between the idea of mining and other research fields and examines some of the previous methods used. Finally, a method is proposed based on two decision tree and machine model algorithms that will improve the results of the idea of mining. The results of the simulation of the proposed method were evaluated and compared with the previous methods. The results show that the proposed method has higher accuracy and speed.

Keywords: Effective Algorithm, Improving the Accuracy, Recommender System, Comments, Classification Techniques, Data Mining

1. INTRODUCTION

Grouping data into tangible categories is one of the basic stages of understanding and learning. In the base state, data grouping systems, depending on whether the data is also shared with the group information, or are read by the observer and unchecked. Cluster analysis is the official study of methods and algorithms for clustering (or grouping) objects based on their inherent characteristics or similarities [1].

Clustering can be used for the following three purposes:

- Finding infrastructure (finding intuition about data, creating hypotheses, detecting abnormalities and identifying outstanding features)
- Natural Data Grouping

- Compression to organize or summarize data

Clustering in various fields of engineering and computer science (machine learning, artificial intelligence, modeling, web mining, data analysis, clustering of texts, image zoning), medical sciences (such as genetics, biology), social sciences, economics Marketing, business) and many other fields of science and engineering.

Web clustering is an important issue for two main reasons. First, if we cluster web pages into a directory, we can easily use it. Auto-classification is important, especially for the World Wide Web, due to the large number of dynamic pages (different times) and the variety of topics, including the problems for categorizing web pages. Second, clustering can improve search and retrieval performance in a document set [2].

Clustering methods are often different in the target function, probabilistic production model, and explorations included. Multiple categories are also proposed based on the input type of the clustering algorithm. Since the variety of clustering methods is very high, for these methods, various categories are presented in the studies. One of the most important categories is categorization in hierarchical and partial hierarchy. Also, clustering algorithms can be classified according to whether they are applied on a data similarity matrix or template matrix including a data property vector [3].

Unfortunately, the clustering of most datasets is a problem. The best clustering algorithm is not in general mode. In other words, there is no clustering algorithm that can comprehensively solve all clustering issues. Each clustering algorithm implicitly or explicitly imposes on the data, and therefore, in cases where there is no match between the model and the data, the algorithm cannot find the appropriate results. The variety of suitable clusters in different applications makes it impossible to create the proper definition of clusters. In the phrase involved, the desirability criterion or clustering purpose can vary from one application to another [4].

In real applications, clusters can have different shapes, sizes and densities. In the last decade, several algorithms have been proposed for the recognition of different clusters, but each of these algorithms has its own weaknesses, and in practice there is not a single algorithm for the detection of cluster structures. The effectiveness of most of the methods introduced on the dataset that has cluster nodes (or overlap) is greatly reduced. While the discovery of clusters with different shapes and sizes usually occurs when the distance between the clusters is relatively low, this state is common in the actual data set [5].

In this research, an antler colony system is used to optimize the clustering algorithm. ACO-based algorithms (ant colony optimization) are able to create the desired clusters without any primary partition and any prior knowledge about how many clusters are needed. Also, ant-based algorithms are suitable for dynamic data and can accommodate the existing clusters and create new clusters or delete existing clusters. Some of the existing methods utilize the versatility and robustness of ant-clustering algorithms and may be able to work with dynamic data. Combined methods combined with standard clustering algorithms are a promising and probable method for online clustering.

2. RELATED WORKS:

The clustering of the structure is in a set of data that has not been categorized. In other words, it can be said that clustering of data is in groups in which members of each group are at a certain angle similar to each other and are not similar to those of other cluster members, or at least have less similar resemblance to cluster members than other members have clusters. Clustering can be used in many applications. Branches such as artificial intelligence, statistics, biology, machine learning, pattern recognition, etc. [6]

In [7], various clustering algorithms are introduced to distinguish different clusters, as discussed below. Of course, some existing algorithms can belong to more than one category. Each category of methods implicitly or explicitly considers models for clusters that usually limit their application to a range of issues. All of these methods, either directly or indirectly, seek to use appropriate definitions for data distributions or clusters.

In [8], the K-Means algorithm is a classical clustering algorithm based on division. This has some inherent problems. First, it covers well-defined values rather than global values. Secondly, the number of clusters should be specified. The selection of the first cluster centers is very important. . In the clustering, the data set is first grouped with the help of evaluation criteria, and then it is grouped into each group as a class. With the help of automated clustering, dense and incomplete regions of our sample space are identified, as well as the distribution pattern of samples and the correlation between the specific traits of each sample. Clustering is also used to explore outside-range data.

In [9], kernel-based methods use the idea of using kernel to find non-linear boundaries between clusters, thus creating the ability to detect clusters with complex shapes.

In [10], most target functions in clustering issues are based on two indices: compression and separation. The compression of data varies in a cluster or between data and cluster centers. And you have to keep it small. Separation is the size of separation of clusters and it is better to be large. If only compression or separation is considered in single-purpose optimization, some of the restrictions may reduce the performance of the optimization. A well-known compression pattern

is known to cause a steady decrease with increasing number of Uniforms.

In [11], a multi-objective evolutionary approach is proposed to optimize data clustering without the need for the number of clusters. The clustering method of fuzzy clusters, in which data related to the degree of membership, includes more information about hard clustering. Fuzzy clustering is possible in many real cases, since the data elements are not exactly assigned to a class. In this paper, the overall mechanism of the FCM-NSGA is presented starting from the Background of the Genetic Algorithm (GAs) and NSGA-II.

In [12] ACO is one of the techniques for approximate optimization. The source of inspiration for the ACO algorithm is the actual behavior of the ants. This is used to solve complex compound optimization problems. It can be used in dynamic applications such as vendor issue, planning, network modeling and routing. Ants live in colonies. Their search behavior is inspired by ACO. They can find the shortest route between their food sources and their nests. Pheromones guide the other ants to reach the food source. Indirect interactions between ants through paths are known as markers.

In [13], a clustering method is proposed for web page clustering with the name of the clone of the growing ant colony (IKACC). This method combines k-means clustering and ant-clustering guided clustering. This method consists of three main modules: 1- Data preprocessing module 2- Clustering module 3- Incremental module.

In [14], the method used does not need to pre-determine the number of clusters, as compared to the growing methods, which are not based on an ant. This method can find the boundaries of clusters without predefined BIOS. In addition, an effective factor in the removal or removal of the entropy is that any action can reduce the entropy of the patch and accelerate clustering. Also, the number of parameters needed to build a clustering module is low.

In [15], a Cluster clustering search algorithm is based on the PSO for combining the strengths of the cuckoo algorithm and PSO. The new solution is based on the PSO. Proposed Hybrid Algorithm with a Dataset Datasheet Testing In most text clustering methods, several preprocessing steps is performed, such as deleting additional words in the data's text. Then, each document is presented again with a frequency vector of the remaining specification in each document.

3. PROPOSED METHOD:

Proposed method is interwoven with the structure of the clustered data display. In fact, this method is provided only for text data, including web pages. Therefore, our only limitation in this method is on the scope of application of the proposed method, and as previously stated, our proposed method is independent of the language content of web pages.

At the beginning of this section, a diagram of the three main steps of the Web page clustering system is presented in the figure, followed by each section in detail.

Figure 1 shows a general outline of the proposed framework. As you can see, this framework consists of three general modules: preprocessing, vector space model, and FCM-ACO based implementation.

3.1 PREPROCESSING

Although this process is not part of the proposed method, but since we intend to provide a thorough and accurate view of the process of clustering web pages using the proposed method, we have explained some key points in this section. It should be noted that the preprocessing process outlined in this section generally has the same format and framework in different natural languages, but its implementation varies depending on the languages. For example, repetitive words, alphabets, and decisions taken against them, while being conceptually identical in different languages, have different equivalents in different languages. In this section, we will refer to the necessary pre-requisites that should be made as the default of the proposed method on the data in the Persian language.

3.2 VECTOR SPACE MODEL

This model is able to efficiently analyze a large collection of files. In fact, this method was initially introduced for data retrieval and indexation, but today it is also used in many text mining methods. In this model, documents are depicted as vectors in the m-dimensional space. In this space, every dimension is a semester. The term semantics is a basic concept, such as a word or phrase. Vector elements correspond to the weight of the term. The document d is represented as d =(x_1, x_2 ... x_n), each x_i shows the importance of the term i in document d. Here we set the similarity based on the distance between the vectors (document with the document). The more vectors are closer to each other, the more similar they are. This similarity is defined by the angle of the two-vector or the cosine-like pattern. The [1] Tf and [2]IDF estimates can be used to assign weight to terms. The Term Frequency (TF[3]) is the count of a single word in a document. The Inverse Document Frequency (IDF[4]) logic is the number of documents that a particular word is viewed in. IDF is used as a weight for TF.

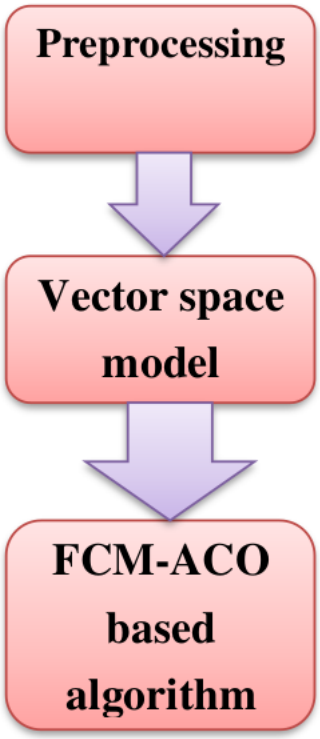


FIGURE 1. Proposed General Framework

The value of this parameter is calculated as follows:

$$IDF_i = \log\left(\frac{N}{n_i}\right) \tag{1}$$

In the above formula, N is the total number of documents in the set and n_i specifies the number of records in which the word i occurred. TF can be calculated directly in any text. Finally, the weight of the word i in the document j is calculated as follows:

$$w_{ij} = TF_{ij} * I\,DF_j \tag{2}$$

After obtaining these weights and displaying each document in the vector space model, we can use the cosine similarity criterion to determine the similarity of the two documents, which is calculated from the following equation:

$$S\left(s_i, s_j\right) = \frac{\sum_{k=1}^{N\sum} w\,k, j_{k,i}}{\sqrt{\sum_{k=1}^{N\sum_k} w\,2_{k,j}}\sqrt{\sum_{k=1}^{N\sum_1} w\,2_{k,i}}} \tag{3}$$

FCM-ACO based algorithm

First of all, we need to point out that the algorithm we offer for clustering web pages, called FCM-ACO based, has several key features that are:

- Privatization of the clustering method for document data only. This means that this method (at least without modifications) can only be used for specific document data;
- No need to determine the number of clusters;
- Modification of the structure of the cluster centers based on the well-known vector-space model and using the colony's colony algorithm;
- The effect of statistical information on data in the initial determination of clusters (rather than completely random selection).

By mentioning these points and drawing attention to these dimensions of the proposed method, the pseudocode of this algorithm has come to be explained and explained to this algorithm.

3.3 PROCEDURE OF FCM-ACO BASED:

```
  Begin
  Initialize
  Sort terms by their weight;
  Calculate the probability of selecting each semester in
the primary cluster centers;
  Create M cluster by genCluster();
  Assigning any clustering to an ant;
  While It is not in the best condition // best cluster
centers
    Repeat
      For each ant do
        Take one step;
        Calculate the position;
        Calculate the fitting position;
        Choose the new position based on pheromones;
        Updates of pheromones based on recent fit and ants
passing;
      End for
  End While
  End
```

4. SIMULATIONS AND COMPARISON OF THE PROPOSED METHOD

In order to evaluate the effectiveness of the proposed method, we consider two different categories of criteria, which correspond to two common approaches in evaluating web clustering methods. The results for each of these criteria are presented in two scenarios. In each section, the effect of the parameter M (number of ant) is determined in the results.

Scenario 1:

Some criteria, such as refreshment, and precision in the field of information retrieval, are defined, which can be interpreted appropriately by them, also the performance of clustering methods. The readout is defined as the proportion of retrieved documentation associated with the total of all relevant documentation. Also, the accuracy is proportional

Recoverable documentation is defined for all recovered data.

In order to calculate accuracy and readout, we must provide the appropriate definitions of the four following categories:

- True Positive (TP)
- True Negative (TN)
- False Positive (FP)
- False Negative (FN)

The definition of TP will be to assign two web pages that are similar to the same cluster. The definition of TN is such that two different web pages are attributed to different clusters. Similarly, FP is equal to the fact that two similar documents are attributed to a different cluster, and FN is defined so that the two pages are uniquely attributed to the same clusters.

Based on the above parameters, Precision (P), (R) Recall and F-score are defined according to the following formulas:

$$P = \frac{TP}{TP + FP} \tag{4}$$

$$R = \frac{TP}{TP + FN} \tag{5}$$

$$F = \frac{2PR}{P + R} \tag{6}$$

Given the popularity of the F-score in the area of Web clustering, the results are reported for this benchmark. Figure 2 shows the performance of the proposed method for the number of ants.

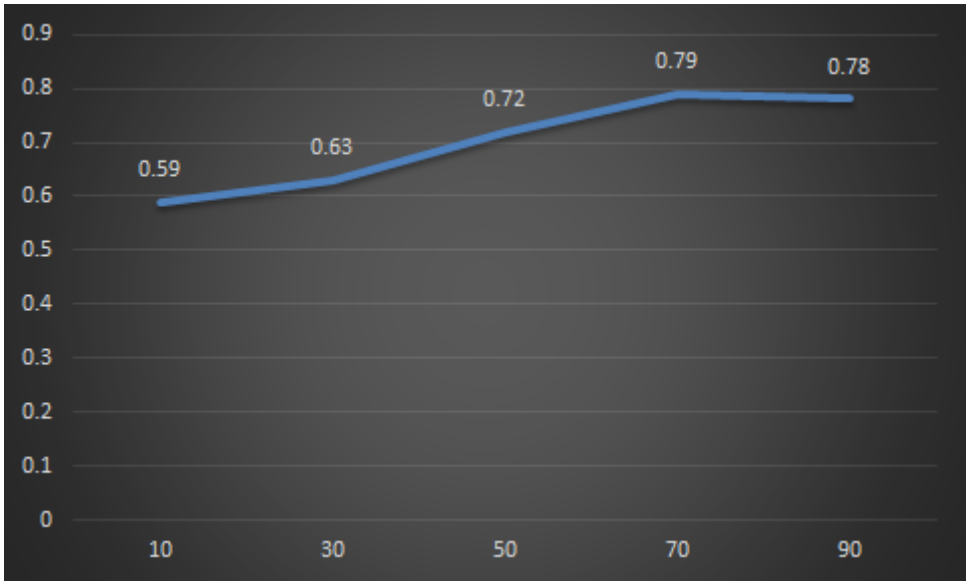
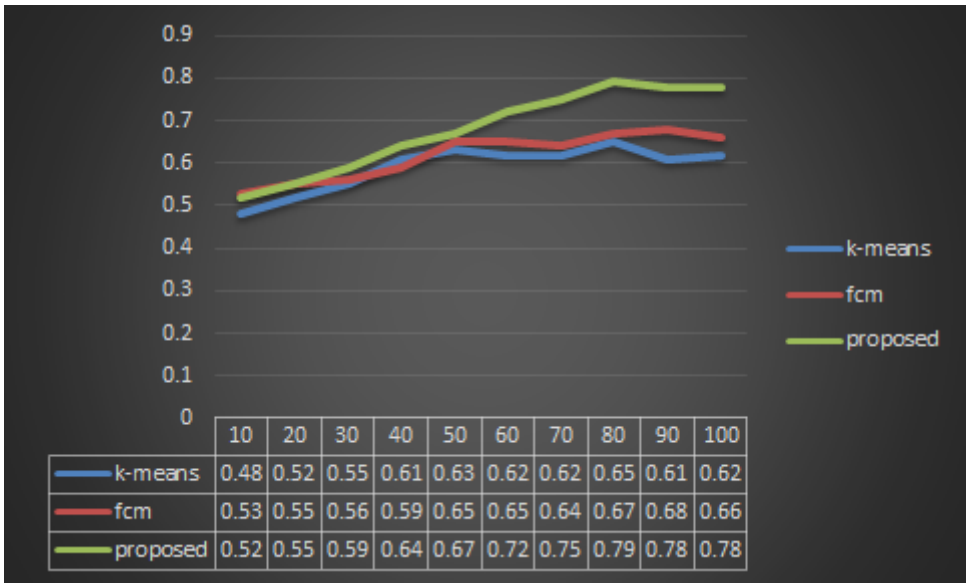


FIGURE 2. The value of F-score for different values of M.

In the figure below, the performance of the proposed method is shown in comparison with the two key methods k-means and FCM for the various volumes of the data set. In fact, in these experiments, various percentages of data sets (from 10% to 100%) are used as data sets.



	10	20	30	40	50	60	70	80	90	100
k-means	0.48	0.52	0.55	0.61	0.63	0.62	0.62	0.65	0.61	0.62
fcm	0.53	0.55	0.56	0.59	0.65	0.65	0.64	0.67	0.68	0.66
proposed	0.52	0.55	0.59	0.64	0.67	0.72	0.75	0.79	0.78	0.78

FIGURE 3. Comparison of the F-score for different methods with the proposed method.

As it is known, in spite of a relatively weaker start with the FCM method, with increasing the size of the dataset (the ratio used from the entire dataset) rapidly, the proposed method provides tangible superiority and stability over the other two methods.

Scenario 2:

We used an index for indexing the indicator in order to examine the proposed method from another point of view. This index can specify the compression of the data in each cluster. The Don's index is defined as:

$$D_{nc} = \min_{i=1,\ldots,nc}\left\{\min_{j=i+1,\ldots,nc}\left(\frac{\operatorname{dist}\left(c_i, c_j\right)}{\max_{k=1,\ldots,nc}\operatorname{diam}\left(c_k\right)}\right)\right\} \tag{7}$$

The number of clusters is nc; the function of the distance between two clusters, which is defined as:

$$\operatorname{dist}\left(c_i, c_j\right) = \min_{x\in c_i, y\in c_j}\operatorname{dis}(x, y) \tag{8}$$

And diam(c) is the diameter of a cluster as defined below:.

$$\operatorname{diam}(c) = \max_{x,y\in c}\operatorname{dist}(x, y) \tag{9}$$

It should be noted that the compression of clusters and their high separation increases the value of this index and vice versa. The results for the three methods K-means, FCM and proposed method are shown in the table below. It is also noteworthy that, in order to determine the number of clusters, clustering was initially carried out by the proposed method and the resultant cluster number was also used for the other two methods.

Table 1. Indices for different methods.

Proposed Method	0.0401
FCM	0.03008
K-means	0.02427

Finally, the efficiency of the proposed method can clearly be derived from the results presented in Table 1. However, this increase, of course, has been at the expense of increasing clustering time, so that the implementation time of the proposed method for the dataset is on average about 2.13 times the FCM method.

5. CONCLUSION

Artificial intelligence techniques that simulate the learning process have been used extensively in a variety of fields, and these applications are increasing dramatically. These techniques have stabilized themselves in a variety of ways, from designing an industrial control system to complex decision support systems. The word learning machine in all these systems is different in different names with different names. This is a process that can seem to tarnish software's or software-driven systems. To implement the proposed method, we are looking for the most efficient tools at the same time. One part of the work that has been done to carry out experiments is in some way related to the separation of dataset files for the desired application and the likelihood of displaying data in the files. For this part of the activities, we have used the Python programming language, which is both academic and academic, and used in the large business of scholars such as Google. . The main reason for this choice is the ease with which files are used in programming languages as well as high-level functions, and there is no other specific technical reason for this choice. In this research, we tried to preserve the least assumptions on the problem space and keep the proposed solution as general as possible. Therefore, the proposed solution has the potential to be used in a wide range of domain domains. With this in mind, this study itself can be a starting point for some of the future research that we will continue to address to them: the use of the proposed method with other evolutionary algorithms To further improve the efficiency of this method in terms of classification accuracy; use of other clustering methods as the basis for clustering; evaluation of the proposed method on the Persian data set.

FUNDING

None

ACKNOWLEDGEMENT

None

CONFLICTS OF INTEREST

The author declares no conflict of interest.

REFERENCES

[1] S. Bandyopadhyay and S. Saha, "A point symmetry-based clustering technique for automatic evolution of clusters," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 11, pp. 1441–1457, 2018.
[2] S. Chakraborty and N. K. Nagwani *Analysis and study of incremental k-means clustering algorithm. High performance architecture and grid computing*, pp. 338–341, 2019.
[3] D. Caro, G. Ducatelle, F. Gambardella, and L. M, "AntHocNet: an ant-based hybrid routing algorithm for mobile ad hoc networks," *In PPSN*, vol. 8, pp. 461–470, 2021.
[4] X. B. Li, Z. W. Wang, K. Peng, and Z. X. Liu, "Ant colony ATTA clustering algorithm of rock mass structural plane in groups," *Journal of Central South University*, vol. 21, no. 2, pp. 709–714, 2014.
[5] W. Min and Y. Siqing, "Improved k-means clustering based on genetic algorithm," *Computer Application and System Modeling (ICCASM), 2010 International Conference on*, vol. 6, pp. 6–636, 2020.
[6] C. W. Tsai, W. L. Chen, and M. C. Chiang, "A modified multiobjective EA-based clustering algorithm with automatic determination of the number of clusters," *2012 IEEE International Conference on*, pp. 2833–2838.
[7] T. İnkaya, S. Kayalıgil, and N. E. Özdemirel, "Ant colony optimization based clustering methodology," *Applied Soft Computing*, vol. 28, pp. 301–311, 2015.
[8] G. Xiangpeng, J. Wang, and S. Liang, "Clustering analysis based on adaptive genetic algorithm for performance assessment," *Control and Decision Conference*, pp. 1682–1686, 2014.
[9] M. Kamarei, G. Kamarei, and Z. Shahsavari, "An Efficient Routing Algorithm to Lifetime Expansion in Wireless Sensor Networks," *Journal of Advances in Computer Research*, vol. 8, no. 1, pp. 107–118, 2017.
[10] H. L. Capitaine and C. Frélicot, "A cluster-validity index combining an overlap measure and a separation measure based on fuzzy-aggregation operators," *IEEE Transactions on Fuzzy Systems*, vol. 19, no. 3, pp. 580–588, 2021.
[11] S. Wikaisuksakul, "A multi-objective genetic algorithm with fuzzy c-means for automatic data clustering," *Applied Soft Computing*, vol. 24, pp. 679–691, 2017.
[12] J. John and R. Pushpalakshmi, "A reliable optimized clustering in MANET using Ant Colony algorithm," *Communications and Signal Processing*, pp. 51–055, 2014.
[13] Y. Boughachiche and N. Kamel, "A New Algorithm for Incremental Web Page Clustering Based on k-Means and Ant Colony Optimization," in *Recent Advances on Soft Computing and Data Mining*, pp. 347–357, Springer, 2019.
[14] C. L. Huang, W. C. Huang, H. Y. Chang, Y. C. Yeh, and C. Y. Tsai, "Hybridization strategies for continuous ant colony optimization and particle swarm optimization applied to data clustering," *Applied Soft Computing*, vol. 13, no. 9, pp. 3864–3872, 2013.
[15] M. M. Zaw and E. E. Mon, "Web document clustering by using pso-based cuckoo search clustering algorithm," *Recent Advances in Swarm Intelligence and Evolutionary Computation*, pp. 263–281, 2015.