Adaptation of Cuckoo Search for Documents Clustering

Walid Mohamed Aly
College of Computing and Information Technology,
Arab Academy for Science, Technology & Maritime Transport
Alexandria, Abu Qir, Egypt

Hany Atef Kelleny
College of Computing and Information Technology,
Arab Academy for Science, Technology & Maritime Transport
Cairo, Masakin Sheraton, Egypt

ABSTRACT

Automatic clustering of unstructured documents has become an essentially indispensible task, especially when dealing with the increasing electronic documents. Automatic clustering of documents involves document designation to a sub-group based on its content. K-means is one of the most popular unsupervised clustering algorithms, though the quality of its results relies heavily on the number of clusters chosen and the right selection of the initial cluster centroids.

Cuckoo search is one of the most recent soft computing intelligent algorithms that can be chosen as an efficient search method in many optimization problems. In this paper, the original Cuckoo Search algorithm is adapted so that it can be applied efficiently to documents clustering problem. Our proposed modification enable Cuckoo search to use dynamic nests so that different values for the number of clusters can be explored, these nests are initialized with different corresponding forgy selection of initial centroids.

During the implementation, these dynamic nests are updated using Lévy flight random walk and evaluated to detect the best nest. The proposed work is implemented and compared to the classical K-means clustering algorithm. The purity measure was used to evaluate the performance. Results show the efficiency of the proposed approach.

Keywords

Cuckoo search, Documents clustering, K-means

1. INTRODUCTION

The majority of today’s organizations encounter a wide range of problems on various levels; essentially they struggle with large volumes of documents flowing every year on a huge scale with all their problems of storing, indexing, retrieval, loss and above all security issues.

The type of “unstructured” data makes up the vast majority of these documents. Handling such amounts of documents posed a daunting task for most of the organizations in spite of the great advances in information technology field.

Unstructured data are simply data that does not fit easily into traditional relational systems. The main problem with unstructured documents is that they have neither indexes, nor key words, or fixed fields to be used in retrievals. In short, documents are created, and in many cases lost in the caves of the organization and in order to restore them back, tremendous efforts should be exerted.

Researchers have noted that today's organizations are faced with exponential growth of unstructured data. Data is changing massively and rapidly, approximately 2.5 exabytes (2.5 million terabytes) of data is created every day [1].

According to Gartner's study in 2010, data growth in 5 years is estimated to be 650% which equates roughly 50% of year over year growth, and 80% of it is unstructured. Moreover, his research indicates that forty exabytes (4.0 x 109) of unique, new information was generated worldwide in 2009 only, which counts to be more than the data produced in the previous 5,000 years.

Similarly, the International Data Corporation (IDC's) newest estimate says that in 2011, the amount of digital information created and replicated in the world was 1.8 zettabytes (1.8 trillion gigabytes or 1.8 quadrillion megabytes), growing 7.9 zettabytes by 2015 [2].

Due to this gigantic growth of electronic unstructured data, the need to develop more systems to dealing with unstructured data became absolutely essential.

Clustering is the act of partitioning unlabeled dataset into groups of similar items/elements. Each cluster comprises of items/elements that are similar among themselves and dissimilar to elements/items of other groups.

Document clustering is a multi-phase unsupervised learning process in which documents are grouped into predefined number of clusters based on their conceptual similarity to each other, without using example documents to establish the conceptual basis for each cluster. This is very useful when dealing with an unknown collection of unstructured text documents. This process commences by performing preprocessing techniques such as tokenization, removing Stop-words or negative dictionary and word stem; then by the operation of indexing the documents, selecting and weighing the terms and then ultimately an internal representation of the documents is created. This representation is frequently found in the form of proximity between pairs of elements. Then the clustering procedure comes next in which the incoming documents are clustered as either hierarchical or partitioned clusters [3].

The goal of our research is adapting of Cuckoo Search as one of recent nature inspired optimization algorithms to solve the process of unstructured documents clustering.

The rest of this paper is organized as follows: Section 2 presents the clustering approaches, Section 3 presents the K-means clustering, and section 4 introduces the Cuckoo search algorithm, section 5 shows the proposed clustering method and section 6 presents results and evaluation for proposed approach, the final section concludes the research.

2. CLUSTERING APPROACHES

Two of the most popular types for documents clustering are subspace and projected clustering. As regards the subspace (overlapping or hierarchical) clustering, it is a clustering approach in which a document might be a member of several clusters in all subspaces, whereas the projected (partitioned or nonhierarchical) clustering, refers to grouping in which each document is assigned to exactly one of non-overlapping clusters, to enhance the evaluation value of clustering.

Clustering algorithms can be classified into hard clustering algorithms and soft clustering algorithms [4].

Hard clustering comprises of two key types algorithms, Flat and hierarchical algorithms, in which K-means is an example of flat clustering algorithms, which aim at partitioning the whole document space into diverse clusters where each cluster includes similar documents. Hierarchical clustering on the other hand uses either a bottom-up or top-down approach to categorize documents with the inception from computing similarity amongst all the documents.

Soft clustering algorithms associate a document to a particular cluster based on the degree of membership of this document to a particular cluster, using membership values to specify the convergence of a document to a cluster. Given that Fuzzy clustering produces a clustering, not a partition, subsequently its documents belong to multiple clusters. Fuzzy C-means is a popular flat soft clustering algorithm, which is rather better than hard K-means.

3. K-MEANS DOCUMENTS CLUSTERING

The K-means algorithm is a popular data-clustering algorithm [5], which is essentially used to group objects based on shared attributes among these objects into k number of groups.

Before the clustering process, each document should be explored to extract its contents of words to be represented as vector of terms.

To build a term vector, the content of a document is analyzed to extract the terms and count their frequencies, whereby preprocessing steps such as stop word removal, stemming should be applied. Then each vector is weighted, typically using a normalized term weighting scheme yielding a set of numerical vectors correspond to the explored documents.

A partition of a set of documents casually finds natural categories among objects by organizing data into clusters such that there is either high intra-cluster similarity or low inter-cluster similarity. It clusters a group of data vectors into a predefined number of clusters by starting with randomly initial cluster centroids and then it keeps on reassigning the data elements to these centroids based on specific similarity coefficient. The ascription procedure would not cease until a convergence criterion is met as shown in Fig.1.

The key input to a clustering algorithm is the Euclidean distance measure as shown in Eq.1.

$$D(x_i, m_k) = \sqrt{\sum_{j=1}^{n} (x_{ij} - m_{kj})^2} \qquad (1)$$

Where

x_i : is i_{th} data object,

m_k : is k$_{\mathrm{th}}$ cluster centroid and

D: is the Euclidean distance.

There are other distance measures that can be used, such as Manhattan distance, Minkowski distance, and Jaccard coefficient [6].

K-means algorithm aims at minimizing the average squared Euclidean Distance of elements from their cluster centers.

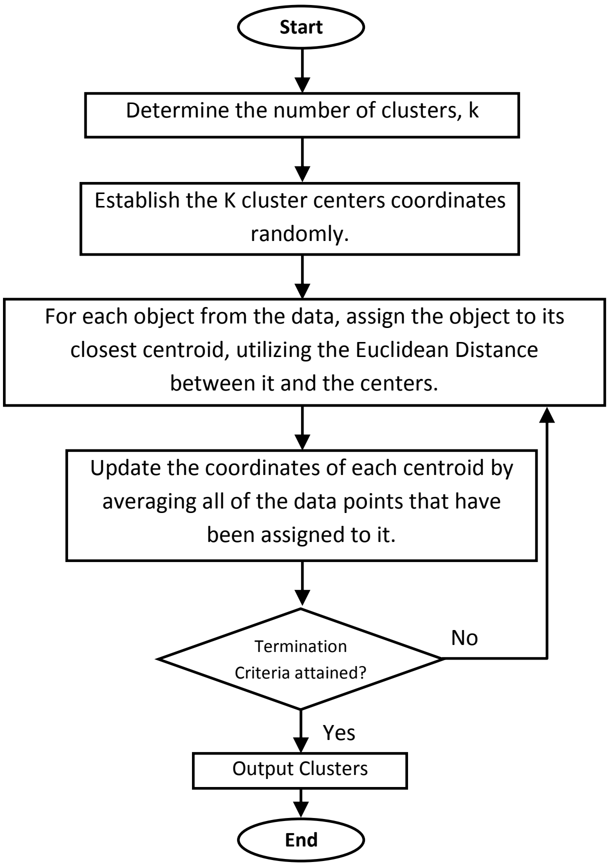
Start
Determine the number of clusters, k
Establish the K cluster centers coordinates randomly.
For each object from the data, assign the object to its closest centroid, utilizing the Euclidean Distance between it and the centers.
Update the coordinates of each centroid by averaging all of the data points that have been assigned to it.
Termination Criteria attained?
No
Yes
Output Clusters
End

Fig 1: K-means Algorithm Flowchart

3.1 K-means Procedure

K-means clustering is a top-down procedure, which comprises of four consecutive steps: initialization, cluster assignment, updating centroids, and finally convergence.

3.1.1 Step1 (Initialization)

Initially, a value for K as the number of clusters is chosen and then centers coordinates are initialized by one of the different methods such as forgy, where some of the data objects (seeds) are selected as initial centroids. Randomly selection is another popular used method to assign the coordinates of these centroids to random values in the search space.

The random method is an appropriate for algorithms such as the k-harmonic means and fuzzy k-means. Meanwhile, the forgy method is recommended for expectation maximization and standard k-means algorithms [7].

3.1.2 Step 2 (Cluster Assignment)

To form the k clusters, each data element is assigned to its nearest center by working out the Euclidean Distance between the element and the centers of clusters.

3.1.3 Step 3 (updating centroids)

To re-estimate the K cluster centers, each cluster center is substituted by the coordinate-wise average of all data points that are closest to it as shown in equation 2.

$$c_j = \frac{1}{n_j} \sum_{\forall d_j \in S_j} d_j \qquad (2)$$

Where d_j denotes the document vectors that belong to cluster S_j; C_j stands for the centroid vector; n_j is the number of document vectors that belong to cluster S_j.

3.1.4 Step 4(convergence)

Repeat step 3 and step 4 until no assignment alteration is achieved or a maximum iteration number is reached.

The main weakness of k-means method is its sensitivity to random initial values of cluster centers [3]. Given that k-means has multiple local optimum and its method cannot ensure its passing these local optimum, thus, convergence to a local optimum could be ensuing if initial position of cluster centers in problem space was selected inappropriately [4]. Subsequently, trying different runs and evaluating their results are typically essential [8].

However, there are exist a number of heuristic methods are used to tune the number of clusters and the initial centroids. Some of these heuristics are presented below.

3.2 Estimating the Number of Clusters

One of the difficult issues in clustering is determining the number of clusters [9], which is denoted by K. K is a good guess based on experience or domain knowledge, however, there are many approximate heuristic methods that can aid in making this decision.

3.2.1 Rule of Thumb

One simple rule of thumb sets the number to:

$$k = \approx \sqrt{\frac{n}{2}} \qquad (3)$$

Where n is the number of objects (data points). This method is not intended to be precisely accurate or reliable for every situation.

3.2.2 The Elbow Method

The elbow method involves graphing the percentage of variance (is the ratio of the between-group variance to the total variance) that is elucidated by the clusters against the number of clusters, and looked for an elbow. The best number of clusters can be determined at the point of this elbow.

To automate and utilize this useful method in the process of documents clustering, the proposed algorithm, the cuckoo search was adapted to be able to handle fluctuating values of K.

4. CUCKOO SEARCH

Xin-she Yang and Suash Deb (2009) introduced the Cuckoo Search Optimization Algorithm. The Cuckoo Search is a search algorithm inspired by the parasitism reproduction behavior of cuckoos [10].

A cuckoo bird lays its egg in randomly chosen nests and relies on other birds for hosting this egg, sometimes these other birds discover the alien egg and demolish it or simply abandon nest.

In aspiration to protect its eggs from detection, the cuckoo bird might replicate the shape, size and color of the host eggs, it might go to the extent of taking an aggressive action by removing other native eggs from the host nest to rise the hatching probability of their own eggs, a hatched cuckoo chick might even throw other eggs away from the nest to improve its feeding share [11].

The concepts of cuckoo reproduction are captured by Cuckoo Search in formulating candidate solutions for an optimization problem as Cuckoo eggs in various nests, the search launches with a fixed number of nests each including a candidate solution to form an initial generation of solution, which evolves from one iteration to another while a fraction of the solution in nests would be eliminated and substituted by new solution to model the concept of the alien egg discovery in a real cuckoo world. This might be drawn-out to the more intricate instances where each nest has multiple eggs signifying a set of solutions.

Cuckoo Search depends on Lévy flight as the random walk, which is used to produce a new mixture (cuckoos) from current solution according to the following equation:

$$cuckoo_i^{t+1} = cuckoo_i^{t} + \propto \oplus\ l\acute{e}vy(\lambda) \qquad (4)$$

Where

$cuckoo_i^{t+1}$: i^{th} Cuckoo at instance $t+1$,

α: step size.

λ: Lévy distribution coefficient

The Lévy flight essentially provides a random walk while the random step length is drawn from a Lévy distribution as follows

$$L\acute{e}vy \sim u = t^{-\lambda}, \quad (1< \lambda \leq 3) \qquad (5)$$

The incidental walk via Lévy flight is more efficient in investigating the search space, as its stride length is much longer on the long term.

Some of the new solutions are produced from the best current solutions by the Lévy walks, which furnish the Cuckoo Search with the capabilities of the local search with the ability of self-improvement as in memetic algorithms. Whereas other new mixtures are produced away from the best current solutions, which reduce the opportunity to be held in local minima, and enhance the versification of the search as in the Tabu search. The employment of Cuckoo Search also ensures elitism, as the best nest would be kept from recurrence to another.

The main advantage of the Cuckoo search is the simplicity of its application since it has fewer parameters that need to be tuned before the inception of the search especially when compared with other techniques, in contrast, PSO needs tuning of mainly three parameters namely, Inertia weight, effect of self-confidence and effect of social impact, where the range of tuning the parameters of PSO noticeably disturb the attribute of search [12]. The crossover rate as well the mutation rate needs to be tuned and various selection methodologies need to be selected.

This cuckoo search algorithm can be simplified and broken down in the following three steps:

1) Each cuckoo bird would lay one egg in each time, and then it would dump its egg in nests that were randomly chosen.

2) The best nests with high quality of eggs would proceed in the following productions.

3) The number of accessible host nests is more or less steady, and the probability of discovering the laid egg by the host bird can be calculated as $p_a \in [0, 1]$. The fraction pa of the n nests is replaced by new nests (with fresh random solutions).

5. THE PROPOSED ADAPTED CUCKOO SEARCH CLUSTERING

The proposed adapted cuckoo search clustering method, involves two successive stages: pre-processing and cuckoo clustering stages. In the initial stage, the pre-processing steps are applied on the incoming documents such as, tokenizing; eliminating stop-words and stemming, subsequently, the documents would be represented as vectors of normalized weights in the space. In the ensuing stage, the clustering stage, the proposed method would apply an adapted cuckoo search based on Lévy flight to cluster those processed documents producing a set of clusters.

Fig. 2 shows the proposed clustering method.

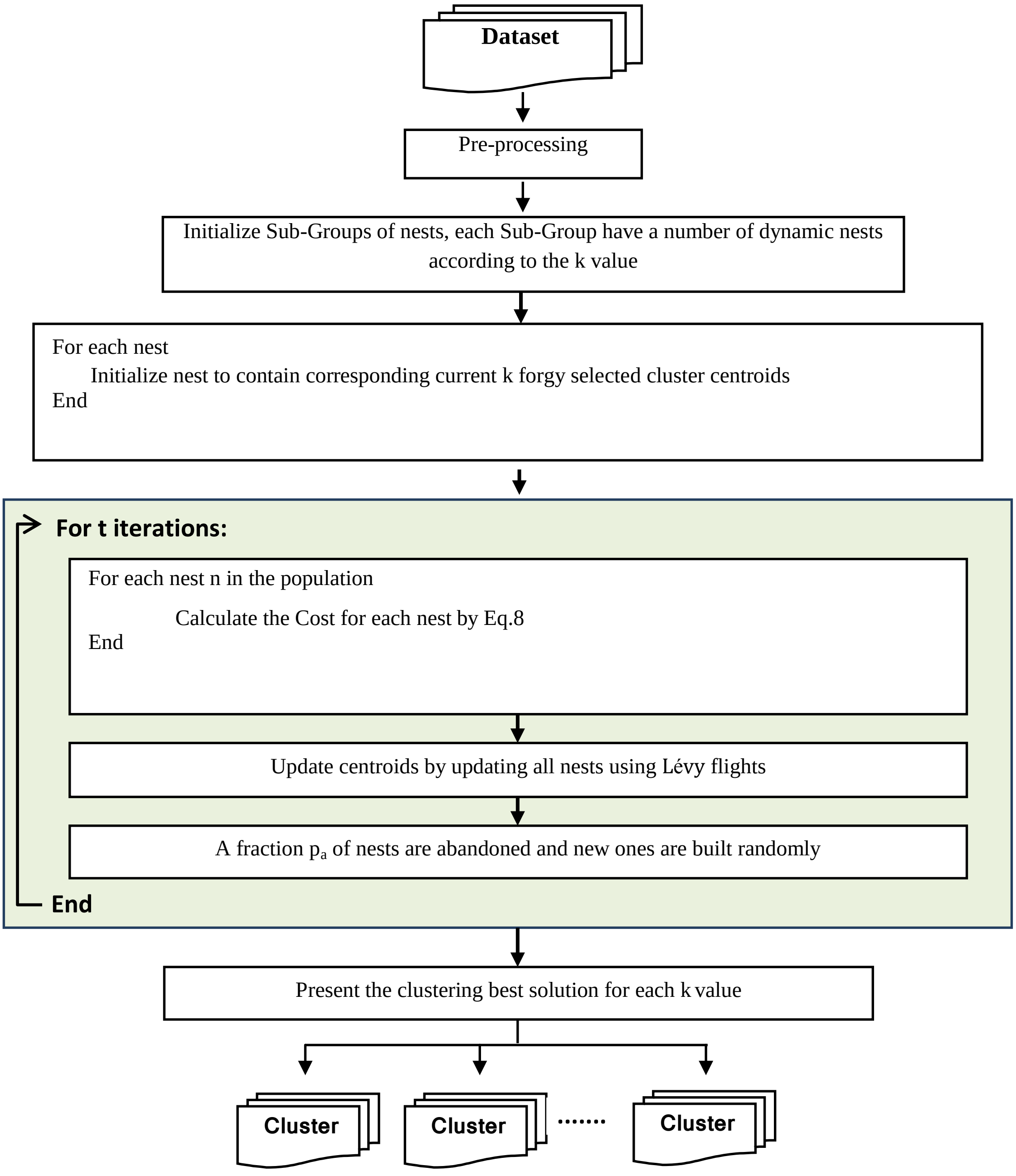
Dataset
Pre-processing
Initialize Sub-Groups of nests, each Sub-Group have a number of dynamic nests according to the k value
For each nest
Initialize nest to contain corresponding current k forgy selected cluster centroids
End
For t iterations:
For each nest n in the population
Calculate the Cost for each nest by Eq.8
End
Update centroids by updating all nests using Lévy flights
A fraction p_a of nests are abandoned and new ones are built randomly
End
Present the clustering best solution for each k value
Cluster
Cluster
....
Cluster

Fig 2: Proposed Clustering Method

5.1 Pre-processing Stage

In the proposed system, the Pre-processing stage involves three steps:-

i. Tokenization,
ii. Stop-words removal, and
iii. Stemming.

A lexical analysis would be performed in order to partition an input stream of characters into a stream of words, producing a list of tokens. To reduce the dimensionality of documents vectors, the proposed system would eliminate the Stop-Words from the incoming documents. Also, The Khoja's stemmer [13] was selected scrupulously based on a comparative study that was conducted among three Arabic morphological analyzers and stemmers, in which Khoja's stemmer was found to achieve the highest accuracy results.

5.1.1 Document Representation

Modern document sets sometimes include millions of documents. To represent them, the proposed system would use the document Vector Space Model (VSM), originated by G. Salton [14] , which represents documents as vectors in a vector space, where each data element (document) di from the corpus D={d1,…,dN } is a represented by a vector of normalized term weights based on the following equation

$$w_{ik} = \frac{tf_{ik}\log(\frac{N}{n_k}+\alpha)}{\sqrt{\sum_{k=1}^{n}(tf_{ik})^2 \times \log^2(\frac{N}{n_k}+\alpha)}} \qquad (6)$$

Where tf_{ik} is the number of occurrences of term i in the document k; and N is the total number of documents in the collection; and n is the total number of documents in the collection, which has the term k [15]:

Because the term may appear in all documents, the α constant is added to the formula, in this case during the calculation of IDF, the log (N/N) = log (1) = 0. The typical value used for α constant is 0.01.

5.1.2 Choosing the Ideal number of Dimensions

Accurate clustering requires a good visual representation for the given set of distances. The Multi-Dimensional Scaling (MDS) algorithm, which is established to this purpose, represents the distances among a set of elements in correct corresponding coordinates in the space according to stress function $\sigma(X)$ shown in Eq. 7.

$$\sigma(X) = \sum_{i<j\le n} w_{ij}\,(d_{ij}(X) - \delta_{ij})^2 \quad (7)$$

Where

$w_{ij} \geq 0$: is a weight for the measurement between a pair of points(i, j).

$d_{ij}(X)$: is the Euclidean distance between i and j.

δ_{ij}: is the ideal distance between the points in the m-dimensional data space.

To choose the ideal number of dimensions, the stress factor should be calculated at different number of dimensions then decide the best degree of dimensionality at the minimum stress. This method has been employed in our proposed clustering method.

5.2 Adapted Cuckoo Clustering Stage

The standard cuckoo search optimization algorithm was adapted and utilized to solve the problem of documents clustering. Adapting the standard cuckoo search was indispensible since it was found as unsuitable for our problem with the fixed sized nests, which is one of the characteristics of standard Cuckoo Search. Supported with initial forgy selection of initial values of centroids, the adapted search is enabled to initialize with a population of nests with a better coverage of search space.

Here, a term is added to each nest structure that would not be enhanced by Lévy flight, yet, it would store the number of clusters (a set of data elements) that this nest represents. So, during the implementation, this bundle of dynamic nests would be categorized into a number of subgroups, which correspond to the initialized different numbers of targeted clusters. Each nest in each subgroup is sized to be appropriate for storing only a number of those targeted cluster centroids.

The initialization of nests ensure that different values of K as number of clusters is represented by a number of nests, nests which represents a certain value of k are considered to be in the same subgroup.

During the search, these dynamic nests are continuously updated utilizing Lévy flight random walk aiming for improvement of each candidate solution.

The following points represent the main procedures included in our proposed approach.

5.2.1 Encoding

Each candidate solution for the clustering problem is encoded as a Cuckoo nest; a nest stores the value of k and the values of corresponding centroids. Fig. 3 shows the typical structure of the nest, which is dynamically sized during its initialization to enable the proper storage of the vector of different values of k cluster centroids. The length of the current nest vector is thus equal to KM +1, where k is the number of current clusters centroids and m is the ideal number of dimensions that were required to achieve an accurate representation of the data objects (Documents) depending on Eq.6.

5.2.2 Cost Calculation

During the search, a cost is calculated for each candidate solution, the less the cost the better the solution. Each data object (document vector) is assigned to its adjacent centroid based on Eq.1, followed by the calculation of the cost function for each nest. A cost function represents the average of the sum of distances between the objects (Document vectors) and their nearest cluster centroids. Accordingly, the cost function for each nest /solution would be calculated by the following equation:

$$Cost = \frac{\sum_{i=1}^{N_c}\{\frac{\sum_{j=1}^{p_i} d(o_i, m_{ij})}{p_i}\}}{N_c} \qquad (8)$$

Where
m_{ij}: denotes the *j^{th}* document, which belongs to cluster *i*.
O_i : is the centroid vector of *i^{th}* cluster.
$d(o_i, m_{ij})$: is the distance between document *m_{ij}* and the cluster centroid *O_i*.
P_i: number of documents that belongs to cluster *C_i*.
N_c: Total number of clusters.

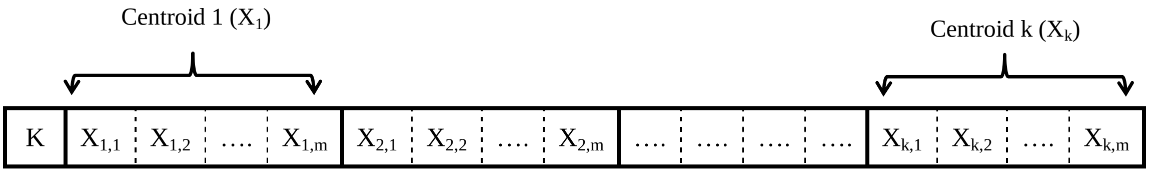

Fig. 3: Nest Structure

5.2.3 Cuckoo Search

On applying the Cuckoo search, on each reiteration, the following steps should be applied.

a-Updating Nests:

All the nests of each subgroup should be updated continuously using Lévy flight as mentioned in Equation 4.

b-Removing of Nests:

On the discovery of alien eggs in its nest, a host bird would either get rid of these eggs or quit its nest altogether in quest of setting up a new one elsewhere. Hence, some nests might be removed which is in line with the host bird's attitude.

To simulate this behavior, a fraction p_a of nests representing candidate solution is removed from iteration and new nests are randomly included to substitute them, this action would prevent the search from being stuck at a local optimum point and ensure a better exploration of search space

c- Elitist selection:

Elitist selection indicate that the best solution is kept unaltered and automatically prorogates from iteration to another except if a better solution is found, this concept is still applied in our approach however in our approach, the best value is each subgroup is kept .

5.3 Setting up Cuckoo Search Algorithm for Documents Clustering

Table 1 shows the parameters used for the applied cuckoo search.

Table 1. Tuning Parameters

Parameter value	value
Number of nests	15
Discovery rate of alien solutions (p_a)	0.25
Lévy exponent (β)	2
Maximum Iteration	100
Number of dimensions	KD +1
Lower bounds for centroids	-1,-1
Upper bounds for centroids	1,1

Where K is the number of cluster centroids and D is the dimensionality degree of the centroids.

6. RESULTS AND EVALUATION

6.1 Dataset

To test and evaluate the proposed approach, we used standard unstructured document set of data, the corpus of contemporary Arabic (CCA) [16]; CCA is composed of 12 different categories. We selected a set of documents form four different categories (politics, tourism and travel, economic and health and medicine) and added all of them in one folder to be used as un-clustered and unlabeled input dataset. CCA is available online from the Website of Leeds University.

6.2 Evaluation Criteria

To evaluate the performance of the proposed method, Purity is used as shown Eq. (9).

$$\text{Purity}(\Omega, \text{C}) = \frac{1}{N}\sum_{k}\max_{j}\left|\omega_k \cap c_j\right| \qquad (9)$$

Where

$\Omega = \left\{ \omega_1, \omega_2, ..., \omega_k \right\}$: is the set of clusters.

$\text{C} = \left\{ c_1, c_2, ..., c_j \right\}$: is the set of classes.

ω_k : is the set of documents in ω_k .

c_j : is the set of documents in c_j .

N: number of clusters

6.3 Results

Table two shows how the results of cost (C) and purity (P) change along with the iterations.

It can be clearly detected that by minimizing the cost function, the purity of clusters was maximized at all different values of k.

Table 2. Performance of proposed adapted Cuckoo Search

	K=3		K=4		K=5	
Iteration No.	**C**	**P**	**C**	**P**	**C**	**P**
20	0.240	0.611	0.192	0.815	0.157	0.755
40	0.236	0.652	0.179	0.821	0.140	0.791
60	0.208	0.723	0.157	0.825	0.134	0.839
80	0.203	0.759	0.150	0.854	0.118	0.843
100	0.179	0.763	0.145	0.874	0.112	0.842

The minimum cost value was achieved at k=5 (0.112) and the purity of clusters at this k value was 0.842. The maximum purity of clusters on the other hand was achieved at k=4 (0.874) although the cost is higher than the cost at k=5. So, k=4 was found to be the best number of clusters as shown in fig 4.

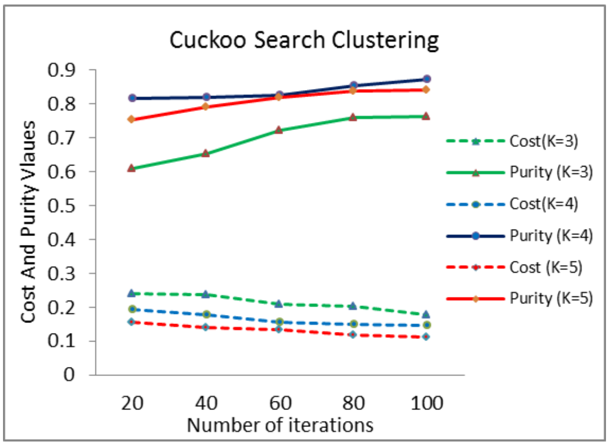

Fig 4: Purity and Cost values at different Iterations

Table 3 illustrates the results of purity measure, which were obtained from the proposed algorithm compared to classical K-means Algorithm in the same dataset, standard K- Mean was initialized using forgy method and was allowed to search until convergence of results occurs.

Table 3. Comparison of results

Clustering method	Purity Measure
K-means clustering algorithm	0.810
Proposed clustering method	0.874

It is clear from the results that the proposed approach acquired a better result than the K-means Clustering and that using Cuckoo Search evolved better rules for document clustering.

7. CONCLUSION

In this research, a new approach for document clustering is presented; this approach relies on Cuckoo Search as a soft computing algorithm to find a solution for the document clustering problem.

The Cuckoo Search optimization Algorithm was adapted for a better performance in solving the problem of concern. Instead of having nests of same sizes, nest with variable sizes was employed to explore the solution space using different number of clusters; moreover, the elitist selection will retain the best solution found for each for the number of clusters, the Forgy method of centroids initialization was used.

The obtained results revealed that the proposed approach has a good performance and reached higher value of clustering purity when compared with classical k-means clustering algorithm.

8. REFERENCES

[1] IBM, July 2012. IBM PowerLinux Big Data Analytics Solutions, USA, Available from: public.dhe.ibm.com [Accessed 24 December 2013].

[2] IDC iView, June 2011.Extracting Value from Chaos, sponsored by EMC. The multimedia content can be viewed at http://www.emc.com/digital_universe [Accessed 20 December 2013].

[3] Krishnamoorthi, M. and Dr. Natarajan A.M., 2013. A Comparative Analysis of Enhanced Artificial Bee Colony Algorithms for Data Clustering. International Conference on Computer Communication and Informatics (ICCCI -2013), Coimbatore, INDIA 4-6 January 2013. IEEE, pp. 1-6.

[4] Singh, V.K., Tiwari, N., and Garg, S., 2011. Document Clustering using K-means, Heuristic K-means and Fuzzy C-means. International Conference on Computational Intelligence and Communication Systems, 2011. IEEE pp. 297-301.

[5] Jensi, R., and Wiselin, G., December 2013. A Survey on Optimization Approaches to Text Document Clustering. IJCSA International Journal on Computational Sciences & Applications, Vol.3, No.6, pp. 31-44.

[6] Rui Tang; Fong, S.; Xin-She Yang; Deb, S., "Integrating nature-inspired optimization algorithms to K-means clustering," 2012 Seventh International Conference on Digital Information Management (ICDIM), 22-24 August 2012. pp.116, 123.

[7] Hamerly, G. and Elkan, C., 2002. Alternatives to the k-means algorithm that find better clusterings. Proceedings of the eleventh international conference on Information and knowledge management (CIKM).

[8] Ahmed MD. E. and Bansal P., 2013. Clustering Technique on Search Engine Dataset using Data Mining Tool. Third International Conference on Advanced Computing & Communication Technologies, Feb. 2013.IEEE, pp.86-87

[9] Agrawal, R.; Phatak, M., 2013. A novel algorithm for automatic document clustering. 3rd International Advanced Computing Conference (IACC), 22-23 Feb. 2013.IEEE, pp.877, 882.

[10] Goel, S.; Sharma, A.; Bedi, P., 2011. Cuckoo Search Clustering Algorithm: A novel strategy of biomimicry. Information and Communication Technologies (WICT) 11-14 Dec. 2011. pp.916, 921.

[11] Xin-She Yang and S. Deb. Cuckoo search via levy flights. In Nature Biologically Inspired Computing, 2009.World Congress on, pages 210–214, 2009.

[12] Aly. W. M. and Sheta, A., 2013. Parameter estimation of nonlinear systems using levy flight cuckoo search. In Max Bramer and Miltos Petridis, editors, Research and Development in Intelligent Systems XXX, pages 443–449. Springer International Publishing, 2013

[13] El-Shishtawy,T., and El-Ghannam ,F.,2012. An accurate arabic root-based lemmatizer for information retrieval purposes. IJCSI International Journal of Computer Science Issues, pp. 58-66.

[14] Salton, G., Wong, A., and Yang, C. S., 1974. A vector space model for automatic indexing.Cornell Univ., Ithaca, NY, USA, Copyright 1975, IEE CU-CSD-74-218.

[15] Zhao,W., Wang ,Y., and Li,D.,2010. A dynamic feature selection method based on combination of ga with k-means. In Seconed International Conference on Industrial Mechatronics and Automation Wuhan, China 30-31 May 2010. pp. 271-274.

[16] Al-Sulaiti,L., Atwell,E.,2006. The design of a corpus of contemporary arabic. International Journal of Corpus Linguistics, vol. 11, pp. 135-171.