International Journal of Innovation and Applied Studies
ISSN 2028-9324 Vol. 4 No. 1 Sep. 2013, pp. 182-188

http://www.issr-journals.org/ijias/

Web Document Clustering Using Cuckoo Search Clustering Algorithm based on Levy Flight

Moe Moe Zaw and Ei Ei Mon

Faculty of Information and Communication Technology,
University of Technology (Yatanarpon Cyber City),
Myanmar

ABSTRACT: The World Wide Web serves as a huge widely distributed global information service center. The tremendous amount of information on the web is improving day by day. So, the process of finding the relevant information on the web is a major challenge in Information Retrieval. This leads the need for the development of new techniques for helping users to effectively navigate, summarize and organize the overwhelmed information. One of the techniques that can play an important role towards the achievement of this objective is web document clustering. This paper aims to develop a clustering algorithm and apply in web document clustering area. The Cuckoo Search Optimization algorithm is a recently developed optimization algorithm based on the obligate behavior of some cuckoo species in combining with the levy flight. In this paper, Cuckoo Search Clustering Algorithm based on levy flight is proposed. This algorithm is the application of Cuckoo Search Optimization algorithm in web document clustering area to locate the optimal centroids of the cluster and to find global solution of the clustering algorithm. For testing the performance of the proposed method, this paper will show the experience result by using the benchmark dataset. The result obtained shows that the Cuckoo Search Clustering algorithm based on Levy Flight performs well in web document clustering.

KEYWORDS: Web Document Clustering, Cuckoo Search, Levy Flight, Clustering Algorithm, Relevant information.

1 INTRODUCTION

The World Wide Web continues to grow rapidly a vast resource of information and services. Powerful search engines have been developed to aid in locating unfamiliar documents by category, contents, or subjects. However, queries often return inconsistent results, with document referrals that meet the search criteria but are of no interest to the user [1].

Clustering's goal is to separate a given group of data items (the data set) into groups called clusters such that items in the same cluster are similar to each other and dissimilar to items in other clusters or to identify distinct groups in a dataset [1]. The results of clustering could then be used to automatically formulate queries and search for other similar documents on the web, or to organize bookmark files, or to construct a user profile. In contrast to the highly structured tabular data upon which most machine learning methods are expected to operate, web and text documents are semi structured. Web documents have well defined structures such as letters, words, sentences, paragraphs, sections, punctuation marks, HTML tags and so forth. Hence, developing improved methods of performing machine learning techniques in this vast amount of non-tabular, semi structured web data is highly desirable.

Text clustering, which has been extensively studied in many scientific disciplines, plays an important role in organizing large amounts of heterogeneous data into a small number of semantically meaningful clusters .In particular, web collection clustering is useful for summarization, organization and navigation of semi-structured Web pages.

Corresponding Author: Moe Moe Zaw (moemoezaw@gmail.com)

One of the best known and most popular clustering algorithms is the k-means algorithm [2]. The algorithm is efficient at clustering large data sets because its computational complexity only grows linearly with the number of data points. However, the algorithm may converge to solutions that are not optimal [3].

PSO is algorithm is presented as document clustering algorithm in [4]. It outperforms K-means clustering algorithm.

In [5], to cluster the web pages, they use the dictionary (standardized) to obtain the context with which a keyword is used and in turn cluster the results based on this context.

A combine approach is proposed to cluster the web pages which first finds the frequent sets and then clusters the documents in [6].

In [7], the Cuckoo Search Clustering Algorithm (CSCA) is proposed. The algorithm is validated on two real time remote sensing satellite- image datasets. In their algorithm, new Cuckoo is calculated by their new equation. In their new Cuckoo solution, the current best solution is considered.

The other new Cuckoo equation for CSCA algorithm is proposed in [8].The new Cuckoo solution is calculated by using both current solution and current best solution.

In [9], the KEA-Means algorithm is proposed and it is used for web page clustering. This algorithm combines key phrase extraction algorithm and k-means algorithm.

This paper describes the application of a new optimization algorithm called Cuckoo Search via Levy Flight [10] to find global solutions to the clustering problem in web document clustering area. The optimization algorithm is based on the obligate brood parasitic behavior of some cuckoo species in combination with the Levy flight behavior of some birds and fruit flies. The algorithm has been successfully applied to different optimization problems including the practical design of steel structure [11] and face recognition [12].

The rest of this paper will present the following. Web document Clustering and Cuckoo Search, the theory background of the algorithm, will be discussed. Then, Cuckoo Search Clustering Algorithm based on Levy Flight will be proposed. Experimental Setup will be presented and results and discussion will also be presented. Then, we will conclude this paper and suggest future work.

2 WEB DOCUMENT CLUSTERING

Articles Web document clustering is widely applicable in areas such as search engines, web mining, information retrieval and topological analysis. Most document clustering methods perform several pre-processing steps including stop words removal and stemming on the document set. Each document is represented by a vector of frequencies of remaining terms within the document. Some document clustering algorithms employ an extra pre-processing step that divides the actual term frequency by the overall frequency of the term in the entire document set.

The text content of a web document provides a lot of information aiding in the clustering of a page. There are many document clustering approaches proposed in the literature. They differ in many parts, such as the types of attributes they use to characterize the documents, the similarity measure used, the representation of the clusters etc. The different approaches can be categorized into *i.texbased, ii.linkbased and iii.hybrid one*. The text-based web document clustering approaches characterize each document according its content, i.e the words (or sometimes phrases) contained in it. The basic idea is that if two documents contain many common words then it is likely that two documents are very similar.

In most clustering algorithms, the dataset to be clustered is represented as a set of vectors $X=\{x_1, x_2, ..., x_n\}$, where the vector xi corresponds to a single object and is called the feature vector. The feature vector should include proper features to represent the object. The web document objects can be represented using the Vector Space Model (VSM) . In this model, the content of a document is formalized as a dot in the multidimensional space and represented by a vector d, such as $d= \{ w_1, w_2, w_n..... \}$, where w_i (i = 1,2,…,n) is the term weight of the term t_i in one document. The term weight value represents the significance of this term in a document. To calculate the term weight, the occurrence frequency of the term within a document and in the entire set of documents must be considered. The most widely used weighting scheme combines the Term Frequency with Inverse Document Frequency (TF-IDF) . The weight of term i in document j is given in Eq (1):

$$w_{j\,i}= tf_{ji}*idf_{ji} = tf_{ji}*log_2(n/df_{ji}) \quad (1)$$

Where tf_{ji} is the number of occurrences of term i in the document j; df_{ji} indicates the term frequency in the collections of documents; and n is the total number of documents in the collection. This weighting scheme discounts the frequent words with little discriminating power.

3 CUCKOO SEARCH

All Cuckoo Search optimization technique is introduced by Yang and Deb recently [10]. Cuckoos have a belligerent reproduction tactic that involves the female laying her fertilized eggs in the nest of another species so that the surrogate parents unwittingly raise her brood [13]. Sometimes the cuckoo's egg in the nest is revealed and the surrogate parents throw it out or dump the nest and start their own brood elsewhere. The cuckoo search optimization algorithm considered various design parameters and constraints, the three main idealized rules on which it is based are as follows[10 [14] :

1) Each cuckoo lays one egg at a time, and dumps its egg in randomly chosen nest;

2) The best nests with high quality of eggs will carry over to the next generations;

3) The number of available host nests is fixed, and the egg laid by a cuckoo is discovered by the host bird with a probability pa€ [0, 1]. In this case, the host bird can either throw the egg away or abandon the nest, and build a completely new nest.

For simplicity, this last assumption can be approximated by the fraction pa of n nests in current iteration which need to be replaced by new nests (with new random solutions) in next iteration. In addition, each nest can represent a set of solutions; CS can thus be extended to the type of metapopulation algorithm.

The number of parameters to be tuned in Cuckoo Search is less than other nature inspired techniques, and thus it is potentially more generic to adapt to a wider class of optimization problems. The technique has been demonstrated successfully on some benchmark functions and is far better, than other approaches including the advanced particle swarm optimization approach [10].

4 CUCKOO SEARCH CLUSTERING ALGORITHM BASED ON LEVY FLIGHT

Cuckoo Search Clustering Algorithm based on levy flight is designed as a clustering algorithm from Cuckoo Search Optimization algorithm to locate the optimal centroids of the cluster. In web document clustering area, it is possible to view the clustering problem as an optimization problem that locates the optimal centroids of the clusters rather than an optimal partition finding problem. This algorithm aims to group a set of input samples (data points) into clusters with similar features. It will work without the knowledge of the class of the input data during the process. In this algorithm, new cuckoo solutions will be moved by using levy flight.

The Cuckoo Search Clustering Algorithm based on Levy Flight is as Fig. 1.

1. Begin

   (Parameter Initialization- no of clusters, no of host nests)
2. Consider NH host nests containing 1 egg (solution) each
3. For each solution of host i
4. Initialize x_i to contain k randomly selected cluster centroids (corresponding to k clusters), as $x_i = (m_{i,1},....,m_{i,j},...m_{i,k})$ where $m_{i,k}$ represents the kth cluster centroid vector of ith cluster centroid vector of i^{th} host.

   End for loop
5. For **t** iterations
6. For each solution of host i of the population
7. For each data document z_p
8. Calculate distance $d(z_p,m_{j,k})$ from all cluster centroids $C_{i,k}$ by using Cosine Similarity Distance eq-2
9. Assign z_p to $C_{i,k}$ by

   $d(z_p,m_{j,k}) = min_{k=1...k} \{ d(z_p,m_{j,k}) \}$

   End for loop in step 7
10. Calculate fitness function $f(x_i)$ for each host nest i by eq-3
11. End for loop in step 6
12. Replace all the nests **except for the best one** by **new Cuckoo eggs produced with levy flight** from their positions
13. A fraction pa of worse nests are abandoned and new ones are built randomly
14. Keep the best solutions (or nests with quality solutions)
15. Find the current best solution

    End for loop in step 5
16. Consider the clustering solution represented by the best solution
17. End

Fig. 1. Cuckoo Search Clustering algorithm based on Levy Flight

$$\cos(m_p,m_j) = m^t_p m_j / |m_p| |m_j| \quad (2)$$

Where: $m^t_p m_j$ = the dot product of the two vector

$$f = \frac{\sum_{i=1}^{N_c} \{ \frac{\sum_{j=1}^{p_i} d(o_i, m_{i,j})}{p_i} \}}{N_c} \quad (3)$$

$m_{i,j}$ = j^{th} document vector which belong to cluster i;

o_i= the centroid vector of i^{th} cluster

$d(o_i,m_{i,j})$= distance between document $m_{i,j}$ and the cluster centroid o_i

p_i = the number of documents which belongs to cluster C_i

N_c= number of clusters

4.1 NEW CUCKOO SOLUTION BASED ON LEVY FLIGHT

The cuckoo laid eggs which correspond to a new solution set. The cuckoo will move from the current position to the new position determined by the levy flight as follows:

$$x_i^{(t+1)} = x_i^{(t)} + \alpha * \text{levy}(\lambda) \quad (4)$$

$$x_i^{(t+1)} = x_i^{(t)} + \alpha * S(x_i^{(t)} - x_{best}^{(t)}).r \quad (5)$$

Where

$x_i^{(t)}$ = current solution
$x_i^{(t+1)}$ = next solution
$x_{best}^{(t)}$ = current best solution
S= random walk based on levy flight
α = step size parameter
r = random number

In Mantegna's algorithm, the step Length can be calsulated by [11].

$$S=\mu/|v|^{1/\beta} \quad (6)$$

Where β is a parameter between [1,2] and considered to be 1.5. μ and v are drawn from normal distribution as

$$\mu \sim N\left(0, \delta_\mu^2\right),\ v \sim N\left(0, \delta_v^2\right) \quad (7)$$

$$\delta_\mu = \left\{\frac{\tau(1+\beta)\sin(\pi\beta/2)}{\tau[(1+\beta)/2]\ \beta\ 2^{(\beta-1)/2}}\right\}^{1/\beta},\ \delta_v = 1 \quad (8)$$

The block diagram of the proposed method is as shown in Fig. 2.

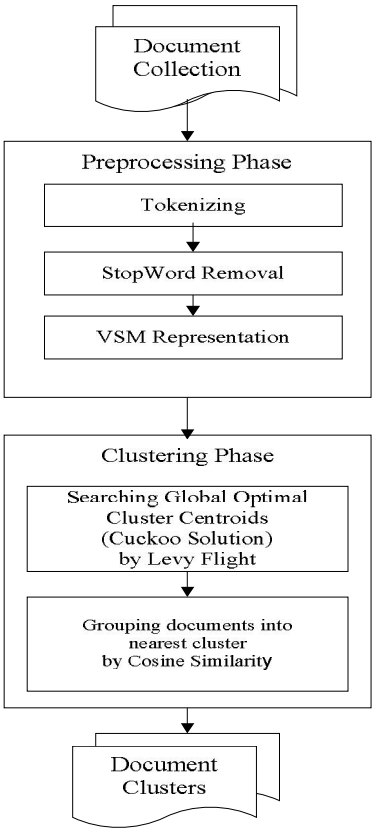

Fig. 2. Block Diagram of the proposed method in web document clustering

In Fig.2, the documents to be clustered must be collected first. The proposed method includes two phases: preprocessing phase and clustering phase. In preprocessing phase, each document will be tokenized and the stop words such as a, an, the etc., will be removed. The remaining words will be represented in Vector Space Model with their TFIDF weight values. In clustering phase, the distance from the center documents to the other documents will be measured by Cosine Similarity measure. The documents to the nearest center will go to this cluster. For next center document selection, the old center will be moved to the new center by Cuckoo Solutions based on levy flight.This clustering process will be performed for a defined number of criteria. The algoithm will finally produce the user-defined number of document clusters.

5 EXPERIENTIAL SET UP

The Cuckoo Search Clustering Algorithm based on levy flight is tested on 7 sector benchmark data set. It is a dataset of collection of web pages of 7 classes. For our testing process, 300 web pages are randomly selected from the dataset and clustered into 3 classes. The algorithm is tested by using Cosine Similarity as distance similarity measure of the two documents. The algorithm executes for 100 iterations and uses 10 cuckoos. The parameter pa is tested for 0.2, 0.25, 0.3 and 0.35. With pa=0.3, the algorithm executes the best fitness value around 50 iterations. So, 0.3 is selected as the pa value of our algorithm. The tested pa values and its cluster quality is as shown in Table.1.

6 RESULT AND DISCUSSION

The fitness equation is also used for the evaluation of the cluster quality. The smaller the cluster quality value, the more compact the clustering solution. The different cluster quality values for the different pa values are as shown in Table 1. The cluster quality values over the number of iterations are as shown in Fig 3.

A famous method for evaluating measure in information retrieval (IR) is F-measure. The cluster results of the system are also evaluated using F-measure. It considers the precision (P), recall (R) and is shown in Eq (9). Eq (10) shows F-measure formula.

$$P = \frac{TP}{TP+FP} \qquad R = \frac{TP}{TP+FN} \tag{9}$$

$$F = \frac{2.P.R}{(P+R)} \tag{10}$$

Table 1. pa values and cluster quality

pa	0.2	0.25	0.3	0.35
Cluster Quality	0.014 ± 0.017	0.012 ± 0.029	0.010 ± 0.014	0.012 ± 0.016

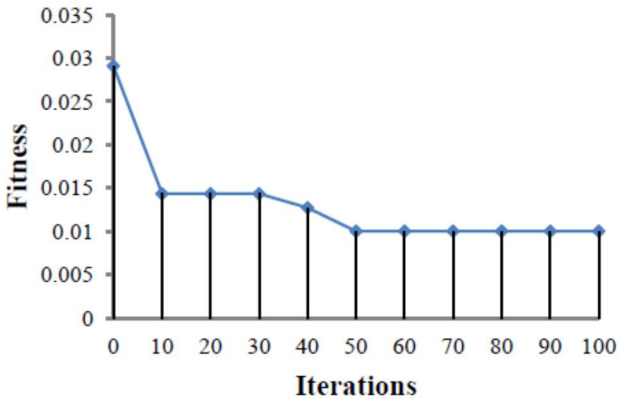

Fig. 3. Performance of Cuckoo Search Clustering Algorithm based on Levy Flight over iteration

Table 2. Precision, Recall and F - measure

Precision	Recall	F-measure
0.711	0.673	0.691

Table.2 illustrates the F-measure of the proposed method. High F-measure shows the high accuracy. The proposed method achieves 0.691 of F-measure in clustering 300 web documents into 3 clusters.

7 CONCLUSION AND FUTURE WORK

Cuckoo Search Clustering Algorithm based on Levy Flight is proposed and applied in web document clustering area. The result shows that the cluster quality and the evaluation measure obtained are good. As our future work, the clustering accuracy can be improved by semantic web document clustering with the help of wordnet, ontology or wekipedia. Our proposed method has been applied in web document clustering area. This Cuckoo Search Clustering Algorithm based on Levy Flight can also be applied to other datasets. And it can also be compared to other swarm intelligence clustering algorithms.

REFERENCES

[1] D. Boley, M. Gini, R. Cross, E. Hong(Sam), K. Hastings, G. Karypis, V. Kumar, B. Mobasher, and J. Moore, "Partitioning-based Clustering for Web Document Categorization", *Decision Support Systems-DSS*, vol. 27, pp. 329-341, 1999.

[2] Jain, A.K. and Dubes, R.C, *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliffs, New Jersey, USA, 1988.

[3] Bottou, L. and Bengio, Y., "Convergence properties of the k-means algorithm", *Advances in Neural Information Processing Systems,* 1995, 7, 585-592.

[4] Xiaohui Cui, Thomas E. Potok, Paul Palathingal, **"**Document Clustering Using Particle Swarm Optimization"*, Swarm Intelligence Symposium, IEEE publication*, 8-10 June 2005.

[5] Jeevan H E, Prashanth P P, Punith Kumar S N, Vinay Hegde, "Web Pages Clustering: A New Approach"*, International Journal Of Innovative Technology & Creative Engineering,* ISSN:2045-8711, vol. 1, no. 4, April 2011.

[6] Rajendra Kumar Roul1, Dr.S.K.Sahay, "An Effictive Web Document Clustering For Information Retrieval", *International Journal of Computer Science and Management Research,* vol. 1, no. 3, p. 481, 2012.

[7] Samiksha Goel, Arpita Sharma, Punam Bedi, "Cuckoo Search Clustering Algorithm: A novel strategy of biomimicry", *World Congress on Information and Communication Technologies, IEEE publication*, 2011.

[8] Moe Moe Zaw, Ei Ei Mon, "Improved Cuckoo Search Clustering Algorithm(ICSCA)", *Proceedings of the 11th International Conference on Computer Applications*, pp. 22-26, 2013.

[9] Swapnali Ware, N.A. Dhawas, "Web Document Clustering Using KEA-Means Algorithm", *International Journal Of Computer Technology & Applications*, vol. 3 (5), pp. 1720-1725, 2012.

[10] Xin-She Yang, Suash Deb, "Cuckoo Search via Levy Flights"*, World Congress on Nature and Biologically Inspired Algorithms, IEEE publication*, pp. 210-214, 2009.

[11] A. Kaveh, T. Bakhshpoori and M. Ashoory, "An Efficient Optimization Procedure Based On Cuckoo Search Algorithm For Practical Design of Steel Structures", *International Journal Of Optimization In Civil Engineering,* 2012; 2(1):1-14.

[12] Vipinkumar Tiwari, "Face Recognition based on Cuckoo Search Algorithm", *Indian Journal of Computer Science and Engineering,* ISSN : 0976-5166, vol. 3, no. 3, Jun-Jul 2012.

[13] X.-S. Yang and S. Deb, "Engineering Optimisation by Cuckoo Search", *International Journal of Mathematical Modelling and Numerical Optimisation,* vol. 1, no. 4, pp. 330-343, 2010.

[14] X.-S. Yang, *Nature-Inspired Metaheuristic Algorithms,* 2nd Edition, Luniver Press, 2010.