A Text-based Web Document Clustering System by Using Improved Cuckoo Search Clustering System Based on Levy Flight

Moe Moe Zaw, and Ei Ei mon

***Abstract*—**The World Wide Web continues to grow rapidly as a vast resource of information and services. The number of web documents on the internet is growing rapidly day by day. So, the process of finding the relevant information on the web is a major challenge in Information Retrieval. The increasing size and the dramatic content of World Wide Web creates a need for automatic categorization of web pages. 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 Improved Cuckoo Search Optimization Algorithm for Global Optimization is a recently developed optimization algorithm to improve the performance of the Cuckoo Search algorithm. In this paper, the Improved Cuckoo Search Clustering Algorithm based on levy flight is proposed to increase the performance of Cuckoo Search Clustering Algorithm based on levy flight. For testing the performance of the proposed method, this paper will show the experience result by using the benchmark dataset. It can be seen that the Improved Cuckoo Search Clustering algorithm based on levy flight outperforms the Cuckoo Search Clustering Algorithm based on levy flight in web document clustering.

***Keywords*—** Web Document Clustering System, Cuckoo Search Clustering System.

I. Introduction

THE increasing size and dynamic content of the World Wide Web has created a need for automated organization of web-pages. Document clusters can provide a structure for organizing large bodies of text for efficient browsing and searching. For this purpose, a web-page is typically represented as a vector consisting of the suitably normalized frequency counts of words or terms. Each document contains only a small percentage of all the words ever used in the web. If we consider each document as a multi-dimensional vector and we try to cluster documents based on their contents of words, the problem differs from classic clustering scenarios in several different ways. Document clustering data is high dimensional, by a highly sparse word-document matrix with positive ordinal attribute values and a significant amount of outliers.

Text clustering has been extensively studied in many scientific disciplines and 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..

One of the best known and most popular clustering algorithms is the k-means algorithm. 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 [1].

PSO is algorithm is presented as document clustering algorithm in [2]. It outperforms K-means clustering algorithm.

In [3], 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 [4].

In [5], 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.

In [6], the Cuckoo Search Clustering Algorithm based on levy flight is proposed. In this algorithm, the new cuckoo solution is generated by levy distribution. The algorithm is applied in web document clustering area and tested with benchmark dataset. The result shows that the algorithm is effective.

The other new Cuckoo equation for CSCA algorithm is proposed in [7].The new Cuckoo solution is calculated by using both current solution and current best solution.

In [8], 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.

In [9], the algorithm Cuckoo Search via Levy Flight is proposed. The algorithm based on the obligate brood parasitic behaviour of some cuckoo species in combination with the L´evy flight behaviour of some birds and fruit flies.

The optimization algorithm is based on the obligate brood parasitic behavior of some cuckoo species in combination with the L´evy flight behaviour of some birds and fruit flies. The algorithm has been successfully applied to different

Moe Moe Zaw is a Ph.D(IT) Student at University of Technology (Yatanarpon Cyber City), Myanmar.

optimization problems including the practical design of steel structure [10] and face recognition [11].

This paper describes the application of a new optimization algorithm called Improved Cuckoo Search Algorithm for Global Optimization [12] to find global solutions to the clustering problem and to enhance the performance of the Cuckoo Search Clustering Algorithm in web document clustering area [6].

The rest of this paper will present the following. Web document Clustering is discussed in next session and then Cuckoo Search Clustering Algorithm will be discussed. Then, Improved 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.

II. Web Document Clustering

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. Textbased, 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.

A. Document Representation

In most clustering algorithms, the dataset to be clustered is represented as a set of vectors $X=\{x_1, x_2, \ldots, x_n\}$.The vector x_i corresponds to a single object and is called the feature vector. The feature vector needs to include proper features to represent the document object. The web document objects can be represented by 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.

B. The Document Similarity Metric

The similarity between two documents is measured in clustering analysis. This algorithm uses the normalized Euclidean distance as the similarity metric of two documents m_p and m_j in the vector space. Eq-1 is the distance measurement formula.

$$d(m_p, m_j) = \sqrt{\sum_{k=1}^{d_m} (m_{pk} - m_{jk})^2 / d_m} \quad (1)$$

where m_p and m_j are two document vectors; d_m is the dimension number of the vector space; m_{pk} and m_{jk} denote the documents m_p and m_j 's weight values in dimension k. This similarity metric is widely used in the text document clustering.

III. Cuckoo Search Clustering Algorithm Based On Levy Flight (CSCA)

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 [6].

The Cuckoo Search Clustering Algorithm based on Levy Flight is as Fig1.

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 $x_i = (m_{i,1},\ldots,m_{i,j},\ldots m_{i,k})$ 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 Euclidean Distance eq-1.
9. Assign z_p to $C_{i,k}$ by
$d(z_p,m_{j,k}) = \min_{k=1\ldots k} \{ d(z_p,m_{j,k}) \}$.
End for loop in step 7
10. Calculate fitness function $f(x_i)$for each host nest i by eq-2.
11. End for loop in step 6
12. Replace all worse nests 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

http://dx.doi.org/10.15242/IIE.E0314053

$$f = \frac{\sum_{i=1}^{N_c}\{\frac{\sum_{j=1}^{p_i} d(o_i, m_{ij})}{p_i}\}}{N_c} \quad (2)$$

$m_{i,j}$ = 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

A. 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:

$$x_i^{(t+1)} = x_i^{(t)} + \alpha * \text{levy}(\lambda) \quad (3)$$

$$x_i^{(t+1)} = x_i^{(t)} + \alpha * S(x_i^{(t)} - x_{best}^{(t)}).r \quad (4)$$

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 calculated by [11].

$$S = \mu/|v|^{1/\beta} \quad (5)$$

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

$$\mu \sim N(0, \delta_\mu^2), \quad v \sim N(0, \delta_v^2) \quad (6)$$

$$\sigma_\mu = \left\{\frac{\Gamma(1+\beta)\sin(\pi\beta/2)}{\Gamma[(1+\beta)/2]\beta 2^{(\beta-1)/2}}\right\}^{1/\beta} \quad (7)$$

IV. Improved Cuckoo Search Clustering Algorithm (ICSCA)

In CSCA, the parameters p_a and α are very important parameters in fine-tuning of solution vectors. They can be potentially used in adjusting convergence rate of the algorithm. If the value of pa is small and the value of α is large, the performance of the algorithm will be poor and lead to increase in number of iterations. If the value of pa is large and the value of α is small, the speed of convergence is high but it may be unable to find the best solutions. So, in the improved Cuckoo Search algorithm, the values of pa and α are dynamically changed with the number of generation.

The Improved Cuckoo Search Algorithm for Global Optimization introduced the following three equations [12]. NI and gn are the total number of iterations and the current iteration, respectively.

$$p_a(gn) = p_{a\,max} - \frac{gn}{NI}(p_{a\,max} - p_{a\,min}) \quad (8)$$

$$\alpha(gn) = \alpha_{max} exp(c.gn) \quad (9)$$

$$c = \frac{1}{NI}\ln[\frac{\alpha_{min}}{\alpha_{max}}] \quad (10)$$

In Improved Cuckoo Search Clustering Algorithm based on levy flight (ICSCA), the new p_a and α values are used to increase the performance of CSCA. According to our knowledge, the dynamically changed values of p_a and α have never been applied in clustering areas. In our system, they are applied in web document clustering area for the first time.

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

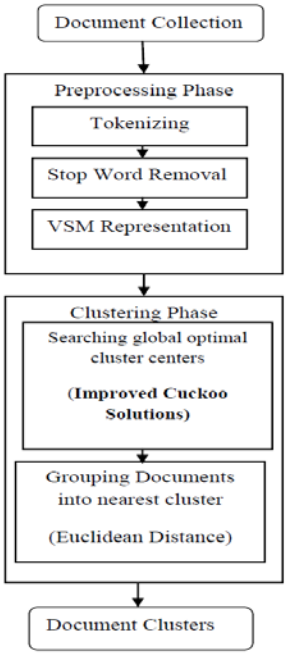

Fig. 2. Block Diagram of the proposed method in Web Document Clustering

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 Euclidean distance similarity measure. The documents to the nearest center will go to this cluster. For next center selection, the old center will be moved to the new center by improved

Cuckoo Solutions based on levy flight. This clustering process will be performed for a defined number of criteria. The algorithm will finally produce the user-defined number of document clusters.

V. Experimental Setup

In CSCA, the parameters p_a and α are very important The Improved 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, 600 web pages are randomly selected from the dataset and clustered into 3 classes. The algorithm is tested by using Euclidean distance as similarity measure of the two documents. The algorithm executes for 100 iterations and uses 10 nests. $p_{a\ min}$ =0.005, $p_{a\ max}$ =1, α_{min} =0.05 α_{max} =0.5 are used. For CSCA, p_a and α are set as 0.25 and 1 respectively.

VI. Results 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 average fitness values of 10 simulations 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 (11). Eq (12) shows F-measure formula.

$$P = \frac{TP}{TP+FP} \qquad R = \frac{TP}{TP+FN} \tag{11}$$

$$F = \frac{2P.R}{(P+R)} \tag{12}$$

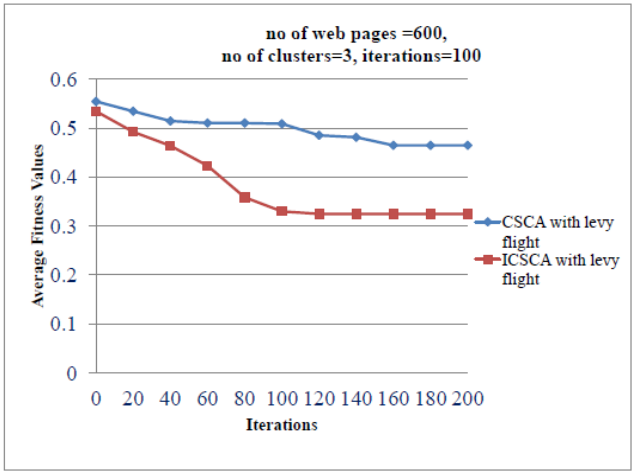

Fig. 3. Performance Comparison of CSCA and ICSCA over iteration

Table[1] illustrates the average F-measure of the proposed method of 100 runs.High F-measure shows the high accuracy. The proposed method achieves 0.646 of F-measure in clustering 600 web documents into 3 clusters while CSCA gains 0.619.

TABLE I
PRECISION, RECALL AND F-MEASURE

Methods	Precision	Recall	F-measure
ICSCA based on Levy flight	0.671	0.623	0.646
CSCA based on levy flight	0.623	0.615	0.619

VII. Conclusion

Improved Cuckoo Search Clustering Algorithm based on Levy Flight is proposed and applied in web document clustering area. Its performance is compared to the performance of Cuckoo Search Clustering algorithm based on levy flight. The result shows that the ICSCA outperforms CSCA. ICSCA performs well in web document clustering area. As our future work, the improved Cuckoo Search Clustering Algorithm based on levy flight can also be applied other clustering areas. And it can also be compared to other swarm intelligence clustering algorithms. The performance can also be improved with excellent feature selection methods and with the help of ontology and wordnet.

References

[1] G. O. Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.

[2] W.-K. Chen, *Linear Networks and Systems* (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[3] H. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1985, ch. 4.

[1] B. Smith, "An approach to graphs of linear forms (Unpublished work Bottou, L. and Bengio, Y,"Convergence properties of the k-means algorithm", Advances in Neural Information Processing Systems, 1995, 7, 585-592.

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

[3] 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

[4] 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

[5] 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.

[6] Moe Moe Zaw and Ei Ei Mon, "Web Document Clustering Using Cuckoo Search Clustering Algorithm based on Levy Flight", International Journal of Innovation And Applied Studies, vol. 4, no 1, pp. 182-188 Sep 2013

[7] 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

[8] Swapnali Ware, N.A.Dhawas Web Document Clustering Using KEA-Means Algorithm", *International Journal Of Computer Technology & Applications*, vol3 (5), pp 1720-1725,2012

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

[10] 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.

[11] 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 .

[12] Ehsan Valian, Shahram Mohanna and Tavakoli,"Improved Cuckoo Search Algorithm for Global Optimization", International Journal of Communications and Information Technology, vol-1,no-1, Dec 2011.