Document Clustering for Knowledge Discovery using Nature-inspired Algorithm

Athraa Jasim Mohammed[1,2], Yuhanis Yusof [1], and Husniza Husni[1]

[1]Universiti Utara Malaysia (UUM), Malaysia, s94734@student.uum.edu.my, {yuhanis, husniza}@uum.edu.my

[2]Information and Communication Technology Center, University of Technology, Baghdad, Iraq

ABSTRACT

As the internet is overload with information, various knowledge based systems are now equipped with data analytics features that facilitate knowledge discovery. This includes the utilization of optimization algorithms that mimics the behavior of insects or animals. This paper presents an experiment on document clustering utilizing the Gravitation Firefly algorithm (GFA). The advantage of GFA is that clustering can be performed without a pre-defined value of *k* clusters. GFA determines the center of clusters by identifying documents with high force. Upon identification of the centers, clusters are created based on cosine similarity measurement. Experimental results demonstrated that GFA utilizing a random positioning of documents outperforms existing clustering algorithm such as Particles Swarm Optimization (PSO) and K-means.

Keywords: Firefly Algorithm, Data Mining, Text clustering, Knowledge Discovery.

1 INTRODUCTION

Information volume in the Internet is growing rapidly and this includes information presented in the form of images and text. Large amount of knowledge is available in textual form and is stored in databases and online sources. In this context, manual analysis and effective discovery of useful information may not be possible. Hence, it would be very useful to provide automatic tools for analyzing the large textual collections. Referring to such needs, data mining tasks such as classification, association analysis and clustering are commonly integrated in the tools.

Clustering is a technique of grouping similar documents into a cluster and dissimilar documents in different clusters (Aggarwal & Reddy, 2014). It is a descriptive task of data mining where the algorithm learns by identifying similarities between items in a collection. Based on literature (Luo, Li & Chung, 2009; Forsati, Mahdavi, Shamsfard & Meybodi, 2013), clustering algorithms can be divided into two main categories; Partitional and Hierarchical. The partitional clustering classifies a collection of documents into a specified number of clusters based on minimizing the distance between documents and center of cluster. The K-means algorithm is a well-known example of partitional clustering as it can easily be implemented. The algorithm operates by dividing objects into groups through the utilization of an error function (Jain, 2010). However, the algorithm may be trapped into local optimum because of the random initialization of centroids (i.e centers of clusters).

On the other hand, the Hierarchical clustering approach constructs a multi-level of clusters (Forsati, Mahdavi, Shamsfard & Meybodi, 2013). It is an efficient method for document clustering in information retrieval as it provides data-view at different levels and organize the document collection in a structured manner. In general, Hierarchical clustering algorithm has two approaches; agglomerative hierarchical clustering and divisive hierarchical clustering. Agglomerative hierarchical clustering merges closest clusters based on dissimilarity matrix while divisive hierarchical clustering splits the cluster into two clusters.

An important issue in clustering is the classification of documents into homogeneous groups. How do we identify initial centroid that minimizes the inter similarity and maximizes the intra similarity? This problem can be stated as an optimization problem. Optimization algorithms find optimal or near-optimal solutions based on an objective function. The objective function can be formulated as a minimum or maximum function depending on the representation of the problem (Rothlaf, 2011). In optimization, the metaheuristic approach is proven to be a successful solution. It can be classified into two categories; single metaheuristic and population metaheuristic solution (Boussaïd, Lepagnot & Siarry, 2013). Single metaheuristic solution initializes with one solution and moves away from it such as implemented in the Simulated Annealing (Kirkpatrick, Gelatt & Vecchi, 1983) and Tabu Search (Glover, 1986). Population metaheuristic solution initializes multi solutions and chooses the best solution based on evaluation of solutions at each iteration such as in Genetic algorithm (Beasley, Bull & Martin, 1993) and nature-inspired algorithms (Bonabeau, Dorigo & Theraulaz, 1999).

The nature-inspired algorithm, also known as Swarm intelligence, is related with the collective behavior of social insect or animal (Rothlaf, 2011). There are many types of Swarm intelligence algorithms such as the Particle Swarm optimization (Cui, Potok & Palathingal, 2005), Ant Colony Optimization (He, Hui & Sim, 2006) and Cuckoo Optimization (Zaw & Mon, 2013). The Firefly algorithm (Yang, 2010) was developed by Xin-She Yang in 2007 at Cambridge University. It has been applied in many disciplines and proven to be successful in image segmentation (Hassanzadeh, Vojodi & Moghadam, 2011) and dispatch problem (Apostolopoulos & Vlachos, 2011). In addition, the FA utilized in numeric data clustering was also a success.

This paper discusses a variant of FA which is known as Gravitation Firefly algorithm (GFA), (Mohammed, Yusof & Husni, 2014) that operates based on random positioning of documents. GFA employs the law of gravity to find force between documents and uses it as the objective function.

The rest of the paper is organized as follows: in section II, we present the standard Firefly algorithm while the Gravitation Firefly Algorithm (GFA) is discussed in section III. Experimental results are discussed in section IV and the conclusion is presented in section V.

II STANDARD FIREFLY ALGORITHM

Firefly algorithm is a swarm intelligent algorithm that is efficient in identifying optimal solution. Firefly algorithm has two important variables; the light intensity and the attractiveness. The light intensity, *I*, of a firefly can be related with objective function *f(x)*. The value of *x* is the location (position) of firefly. Every location has different value of light intensity. The objective function can be maximized or minimized depending on the problem. The attractiveness, *β*, is related with light intensity. Relatively, it means that when two fireflies are attracted between each other, the highest intensity will attract the lower intensity and the value of *β* changes based on the distance between two fireflies. The attractiveness, β, formula is shown in Eq. (1) (Yang, 2010).

$$\beta = \beta_0 exp^{(-\gamma r_{ij}^2)} \tag{1}$$

Where, β_0 is the attractiveness when the distance r has value 0. Y is the absorption coefficient value between (0-1).

The movement of one firefly *i* to another firefly *j* is determined based on Eq. (2).

$$x^i = x^i + \beta * \left(x^j - x^i\right) + \alpha\varepsilon_i \tag{2}$$

Where, x_i is the position of first firefly; x_j is the position of second firefly. ε_i refers to random numbers between 0 and 1.

The pseudo-code of standard Firefly Algorithm is shown in Figure 1.

```
Objective function f(x), x=(x1, ..., xn)T
Generate Initial population of firefly randomly
  xi ( i=1, 2, .., n)
Light Intensity I at xi is determine by f(xi )
Define light absorption coefficient γ
While (t < Max Generation)
For i=1 to N (N all fireflies)
For j=1 to N
If (Ii < Ij) { Move firefly i towards j; end if
Vary attractiveness with distance r via exp[-γr]
 Evaluate new solutions and update light
  intensity
 End For j
 End For i
 Rank the fireflies and find the current global
  best g*
 End while
 Postprocess results and visualization
```

Figure 1. Pseudo-code of standard Firefly Algorithm (Yang, 2010)

III GRAVITATION FIREFLY ALGORITHM

Gravitation Firefly Algorithm (GFA) (Mohammed, Yusof & Husni, 2014) is an approach in document clustering that utilizes law of gravity as the objective function. GFA employs the law of gravity to find force between documents and uses as maximizing objective function. The objective function is based on similarity between documents and the distance. The distance between documents is calculated using position of document in search space. The Newton's law of gravity is stated that "*Every point mass attracts every single other point mass by a force pointing along the line intersecting both points. The force is proportional to the product of the two masses and inversely proportional to the square of the distance between them.*" (Rashedi, Nezamabadi-pour & Saryazdi, 2009). The Newton's law of gravity formula is as shown in Eq. (3).

$$F = G\frac{M_1 * M_2}{R^2} \tag{3}$$

Where: F is the force between two masses, G is the gravitational constant, M_1 is the first mass, M_2 is the second mass, R is the distance between two masses.

GFA identifies document having the highest value of force as center of clusters (i.e centroid). Each document is represented by a single firefly. The force between documents is computed using Eq. (3) where *G* is represented by the similarity between two objects, M_i and M_j are the mass of two documents (we supposed that the mass is the summation of all terms weight in a document) (Mohammed, Yusof & Husni, 2014) and *R* is the distance between two positions of documents in asearch space and is calculated using the Euclidean distance. In this paper, the position of each firefly (document) in the search space is represented by coordinates (x,y).

IV EXPERMINTAL RESULTS

A standard benchmark text dataset, called 20 Newsgroups (20 Newsgroup Dataset, 2006), is utilized in evaluating the proposed GFA. We choose 300 documents from 3 classes. The description of the collection is provided in Table1.

Table 1. The 20 newsgroups Dataset.

Dataset Topics	No. of Documents	Total No. of Classes	No. of Terms
Comp.sys.mac.hardware	100	3	2215
Rec.sport.baseball	100		
Sci.electronic	100		

The initial position of GFA is illustrated in Figure 2, where *x* is a random value in the range (1-300) and *y* is fixed at 0.5.

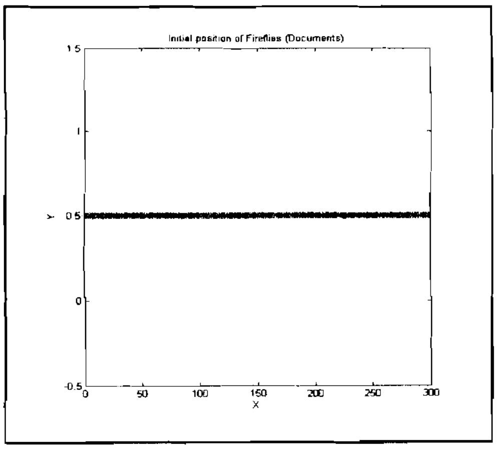

Figure 2. Initial position of Fireflies (Documents)

The position of documents upon completing the 20th iteration is illustrated in Figure 3 until Figure 7.

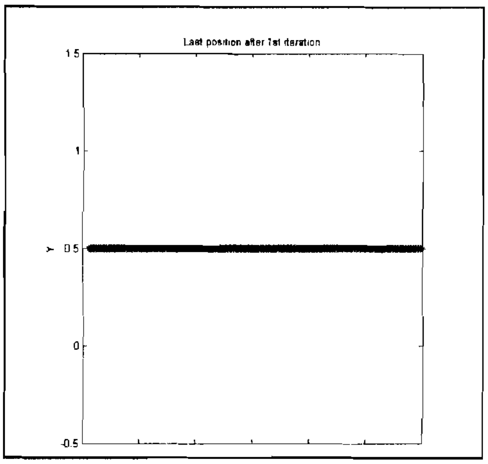

Figure 3. Fireflies (Documents) Position After 1st Iteration

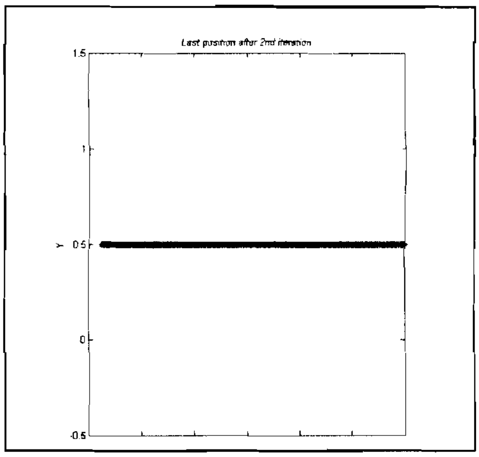

Figure 4. Fireflies (Documents) Position After 2nd Iteration

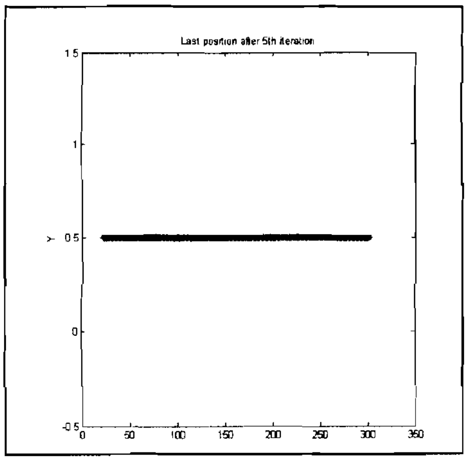

Figure 5. Fireflies (Documents) Position After 5th Iteration

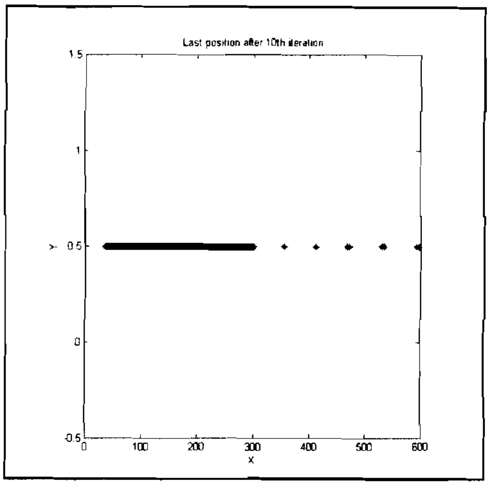

Figure 6. Fireflies (Documents) Position After 10th Iteration

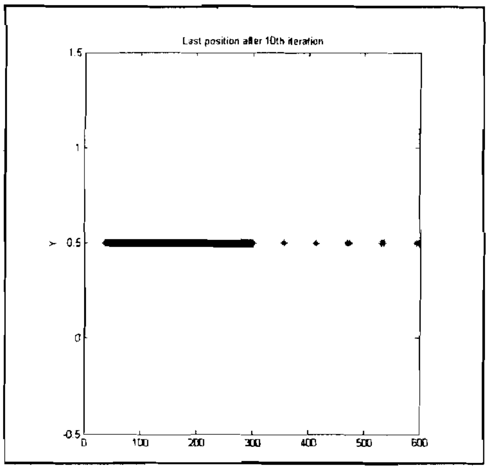

Figure 7. Fireflies (Documents) Position After 20th *Iteration*

The algorithm is evaluated using quality performance metrics that includes average distance between each centers and documents (ADDC), Purity, F-measure and Entropy (Forsati, Mahdavi, Shamsfard & Meybodi, 2013). The GFA performance is compared againts two algorithms; K-Means and Particle Swarm Optimizatio (Cui, Potok & Palathingal, 2005). Figure 8 shows the ADDC of GFA, PSO and K-means. From the Figure, we can illustrate the convergence behaviors of the three techniques; GFA (Mohammed, Yusof & Husni, 2014), PSO (Cui, Potok & Palathingal, 2005) and K-means. The graph illustrated in Figure 8 shows that GFA has the smallest value of ADDC. A smaller value of ADDC is indicates a better cluster and it satisfies the optimization constrains (Forsati, Mahdavi, Shamsfard & Meybodi, 2013).

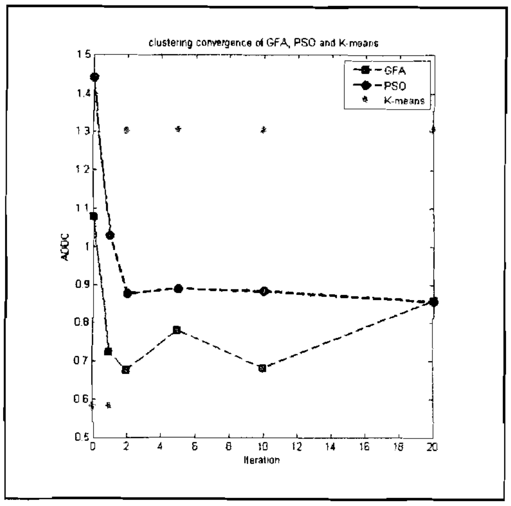

Figure 8. The ADDC of GFA, PSO and K-means

Figure 9 shows the Purity of GFA, PSO and K-means. It shows that the purity of GFA increases starting from iteration 10 until 20 and it produces the highest value 0.72 in iteration 20 while PSO generates the smallest value.

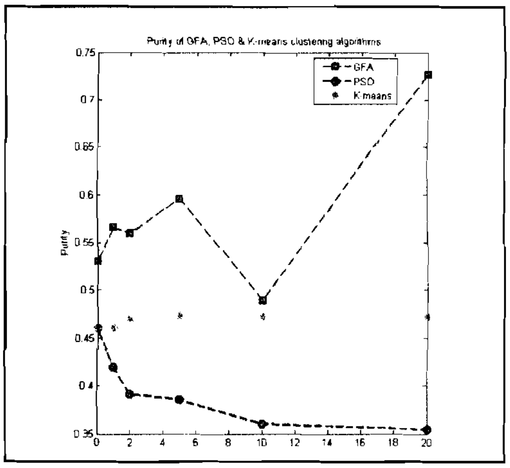

Figure 9. The Purity of GFA, PSO and K-means

In Figure 10, the F-measure of GFA, PSO and K-means is presented. It shows that performance of GFA increases as the number of iteration. The highest F-measure value is 0.593, obtained in iteration 20 while PSO obtains 0.541 at iteration 5. On the other hand, the F-measure performance of K-means technique is 0.473 in iteration 5.

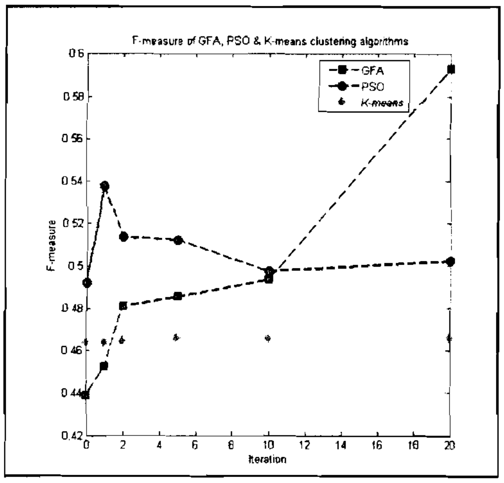

Figure 10. The F-measure of GFA, PSO and K-means

Figure 11 shows the Entropy obtained by GFA, PSO and K-means. The entropy of GFA is the best as it generates the smallest value compared to the ones by PSO and K-means.

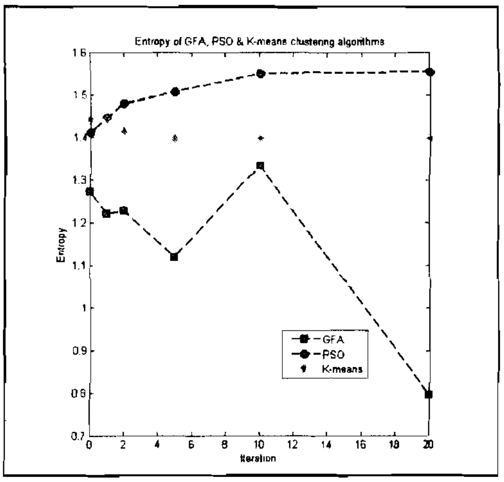

Figure 11. The Entropy of GFA, PSO and K-means

V CONCLUSION

The World-Wide Web provides users with access to abundance of information, but it becomes increasingly difficult to discover relevant pieces of information. Research in knowledge discovery tries to address this problem by applying techniques from data mining and machine learning to Web documents. In this paper, we present a method for knowledge discovery that would benefit the organization of text collections. The undertaken approach is based on nature-inspired algorithm which is the Gravitation Firefly Algorithm (GFA). The proposed GFA utilizes random positioning of documents in grouping text documents automatically. GFA determines the center of clusters based on the gravitation law. Furthermore, the positioning of documents, prior to clustering, is undertaken based on random initialization. Empirical study indicates that the proposed GFA overcomes commonly used clustering techniques such as Particles Swarm Optimization and K-means.

Mohammed, A. J., Yusof, Y. & Husni, H. (2014). A Newton's Universal Gravitation Inspired Firefly Algorithm for Document Clustering. *Proceedings of Advanced in Computer Science and its Applications*, v. 279, Lecture Notes in Electrical Engineering,

ACKNOWLEDGEMENT

Authors would like to thank the Ministry of Education (MOE) for providing the financial support under the Fundamental Research Grant (s/o 12894).

REFERENCES

Apostolopoulos, T., & Vlachos, A. (2011). Application of the Firefly Algorithm for Solving the Economic Emissions Load Dispatch Problem. International Journal of Combinatorics, Volume 2011 (2011), Article ID 523806, 23 pages.

Beasley, D., Bull, D. R., & Martin, R. R. (1993). An Overview of Genetic Algorithms : Part 1, Fundamentals. University Computing, 15(2), 58-69.

Bonabeau, E., Dorigo, M., & Theraulaz, G. (1999). Swarm Intelligence: From Natural to Artificial Systems: New York, NY: Oxford University Press, Santa Fe Institute Studies in the Sciences of Complexity.

Boussaïd, I., Lepagnot, J., & Siarry, P. (2013). A survey on optimization metaheuristics. Elsevier, Information Sciences, 237, 82-117.

Cui, X., Potok, T. E., & Palathingal, P. (2005, 8-10 June 2005). Document Clustering using Particle Swarm Optimization. Paper presented at the Proceedings 2005 IEEE Swarm Intelligence Symposium, SIS 2005.

Forsati, R., Mahdavi, M., Shamsfard, M., & Meybodi, M. R. (2013). Efficient stochastic algorithms for document clustering. Elsevier, Information Sciences, 220, 269–291.

Glover, F. (1986). Future paths for integer programming and links to artificial intelligence. Computers and Operations Research, 13(No.5), 533-549.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by Simulated Annealing. New Series, 220(No. 4598), 671-680.

Luo, C., Li, Y., & Chung, S. M. (2009). Text document clustering based on neighboors. Elsevier, Data & knowledge Engineering, 68(11), 1271-1288.

Hassanzadeh, T., Vojodi, H., & Moghadam, A. M. E. (2011). An Image Segmentation Approach Based on Maximum Variance Intra-Cluster Method and Firefly Algorithm. Paper presented at the Seventh International Conference on Natural Computation (ICNC), Shanghai.

He, Y., Hui, S. C., & Sim, Y. (2006). A novel ant-based clustering approach document clustering. In H. Tou Ng, M. K. Leong, M. Y. Kan & D. Ji (Eds.), Information Retrieval Technology (Vol. 4182, pp. 537-544): Springer Berlin Heidelberg.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Elsevier, Pattern Recognition Letters, 31(8), 651–666.

Jeong, H.Y., Yen, N. Y., Park, J.J. (Jong Hyuk), Springer Berlin Heidelberg, pp. 1259-1264.

Mohammed, A. J., Yusof, Y. & Husni, H. (2014). Weight-Based Firefly Algorithm for Document Clustering. *Proceedings of the*

First International Conference on Advanced Data and Information Engineering (DaEng-2013), v. 285, Lecture Notes in Electrical Engineering, Herawan, T., Deris, M. M., Abawajy, J., Springer Berlin Heidelberg, pp. 259-266.

Rashedi, E., Nezamabadi-pour, H., & Saryazdi, S. (2009). GSA: A Gravitational Search Algorithm. Elsevier, Information Sciences, 179(13), 2232-2248.

Rothlauf, F. (2011). Design of Modern Heuristics Principles and Application: Springer-Verlag Berlin Heidelberg.

Yang, X. S. (2010). Nature-inspired metaheuristic algorithms 2nd edition. United Kingdom: Luniver press.

Zaw, M. M., & Mon, E. E. (2013). Web Document Clustering using Cuckoo Search Clustering Algorithm based on Levy Flight. International Journal of Innovation and Applied Studies, 4(no.1), 182-188.

Aggarwal, C. C., & Reddy, C. K. (2014). Data clustering algorithm and applications: CRC press, Taylor and Francis Group.