Journal of Engineering and Applied Sciences 11 (3): 522-527, 2016
ISSN: 1816-949X

Integrated Bisect K-Means and Firefly Algorithm for Hierarchical Text Clustering

[1, 2]Athraa Jasim Mohammed, [1]Yuhanis Yusof and [1]Husniza Husni
[1]School of Computing, College of Arts and Sciences, Universiti Utara Malaysia,
06010 Sintok, Kedah, Malaysia
[2]Information and Communication Technology Center, University of Technology,
Baghdad, Iraq

Abstract: Hierarchical text clustering plays a significant role in systematically browsing, summarizing and organizing documents into structure manner. However, the Bisect K-means which is a well-known hierarchical clustering algorithm is only able to generate local optimal solutions due to the employment of K-means as part of its process. In this study, we propose to replace the K-means with firefly algorithm, hence producing a Bisect FA for hierarchical clustering. At each level of the proposed Bisect FA, firefly algorithm works to produce the best clusters. For evaluation purposes, we performed experiments on 20 newsgroups dataset that is commonly used in text clustering studies. The results demonstrate that Bisect FA obtains more accurate and compact clustering than Bisect K-means, K-means and C-firefly algorithms. Such a result indicates that the proposed Bisect FA is a competitive algorithm for unsupervised learning.

Key words: Hierarchical text clustering, firefly algorithm, bisect K-means, divisive clustering, documents

INTRODUCTION

Traditional methods in clustering can be divided into five types; partitional clustering, hierarchical clustering, density clustering, model based clustering and grid based clustering (Han *et al*. (2011); Zhang *et al*., 2013). This paper focuses on the partitional clustering and hierarchical clustering of text documents. Partitional clustering groups objects into specific number of k clusters based on some criterion (e.g., Sum of Squared Error (SSE)). The K-means (Jain, 2010) and Fuzzy C-Means (FCM) (Zhong *et al*., 2010) are the two mostly used traditional clustering algorithms. This is due to their simplicity, efficiency and speedy convergence. The difference between these two algorithms is that K-means is a hard clustering while FCM is a soft clustering (Aliguliyev, 2009).

Hard clustering requires every object is assigned to only one cluster while soft clustering allows various membership degrees. The steps in K-means clustering (Ain, 2010) are:

Steps in K-means (Ain, 2010):

- Step 1; randomly choose k cluster centroids
- Step 2; assign each object to closest centroid
- Step 3; recalculate the centroids
- Step 4; repeat step1 and step 2 until stopping condition is reached

To start, K-means randomly identify a number of k centroids and assigns objects to their closest centroid by minimizing the Sum of Squared Error (SSE). Then, K-means updates the centroid of each cluster by calculating the mean of all objects that belong to the specific cluster. K-means will stop its execution once a predefined number of iterations have been exceeded or a stagnant error rate is obtained (Jain, 2010; Rokach and Maimon, 2005). Clustering using K-means is sensitive to the initial centroids selection, hence may result in a local optima problem. Another drawback of K-means is its dependency on the number of k clusters (Cui *et al*., 2006). With that being said, researchers have moved to FCM which is a variant of K-means that overcomes the local minima. Nevertheless, it still has the problem with the design of membership function (Rokach and Maimon, 2005).

Hierarchical clustering constructs multi-level clusters by recursively grouping the objects using either two directions; top down (divisive methods) or bottom up (agglomerative methods) (Forsati *et al*., 2013). A divisive clustering method operates by dividing all objects that belong to one cluster into specific number of clusters. The

Corresponding Author: Athraa Jasim Mohammed, School of Computing, College of Arts and Sciences, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia

Bisect K-means (Murugesan and Zhang, 2011b; Kashef and Kamel, 2009). is a well-known divisive hierarchical clustering and is a variant of K-means. In this algorithm, at each level of constructing a hierarchy, Bisect K-means selects one cluster, C (initially C represents the whole dataset) and classifies the objects into two partitions (C1 and C2) by randomly choosing two centers and assigning objects to the closest centers (using K-means algorithm).

This process continues until it reaches the stopping condition as either number of iterations or specific number of clusters. At each step of classifying, the chosen cluster is tested by some criteria: minimum intra similarity; the larger cluster size (means cluster includes large number of objects) or size of cluster and similarity (Murugesan and Zhang, 2011a; Kashef and Kamel, 2009).

Shows the steps in Bisect K-means are:

- Step 1; randomly choose two cluster centroids
- Step 2; cluster using K-means
- Step 3; if not reach number of clusters, choose the cluster that has smallest intra similarity for further process
- Step 4; repeat step1 until reach number of clusters

Background: Bisect K-means requires a refinement step to re-cluster the resulting solutions at each level of constructed tree. This drawback attracts researchers to combine Bisect K-means with K-means. In the work of (Kashef and Kamel, 2009, 2010). The clustering solution of Bisect K-means and K-means at each level cooperated between them by cooperative and merging matrices. Further, the Un-weighted Pair Group Method with Arithmetic Mean (UPGMA) (a type of agglomerative clustering) merges the obtained clusters from Bisect K-means (where, Bisect K-means generates clusters larger than k) until it reaches the k number of clusters (Kashef and Kamel, 2009; Murugesan and Zhang, 2011a, b). In general, Bisect K-means uses k-means at each level of tree construction. Nevertheless, K-means is sensitive to the initial centroids selection, hence causing a local optima problem.

Existing studies show that optimization algorithm is an alternative in solving local optima problem. Generally, the goal of clustering is to achieve high similarity among objects in a cluster and less similarity between clusters. Such a situation can be represented as an optimization problem (Banati and Bajaj, 2013). Optimization identifies the best solution (optimal or near optimal solution) from a set of available solutions using an objective function (can be formulated as minimum or maximum). The design of an objective function is based on the problem in-hand (Rothlauf, 2011). Recently, meta-heuristic approach has proven to be a success in finding the best solution (Kirkpatrick *et al*., 1983; Cui *et al*., 2006). Existing meta-heuristic approach can be divided into two groups; single solution and population solution (Boussaid *et al*., 2013). Single solution meta-heuristic starts with a single solution and tries to enhance it while population meta-heuristic solution starts with a set of solutions and evaluate them to choose the best one. Simulated Annealing (Kirkpatrick *et al*., 1983) and Tabu Search (Glover, 1986) are examples of single solution meta-heuristic while Genetic algorithm (Beasley *et al*., 1993). Evolutionary programming (Fogel, 1994). Differential Evolution (Rokach and Maimon, 2005) and nature-inspired algorithms (Fogel, 1994) are types of population meta-heuristic solution.

Nature-inspired (also called as Swarm intelligence) algorithms includes studies on social insects or animal behaviors in the nature and mimics these behaviors to solve problems faced by humans (Rothlauf, 2011). Swarm intelligence algorithms include the Particle Swarm Optimization (Kennedy and Eberhart, 1995). that studies behavior of the flock and foraging, Ant Colony Optimization (He *et al*., 2006). that imitates the behavior of ants and Cuckoo Optimization (Zaw and Mon, 2013) that mimics the cuckoo behavior. In the work reported by Tang *et al*. (2012a) an integration of nature inspired optimization with K-means for clustering is presented. The optimization methods include the Wolf (Tang *et al*., 2012b), Firefly (Yang, 2010). Cuckoo (Yang and Deb, 2009). Bat (Yang, 2010) and Ant Dorigo such an integration is proposed to guide the searching for global optima and speed up the convergence.

The Firefly Algorithm (FA) (Banati and Bajaj, 2013; Yang, 2010) is an algorithm proposed by Xin-Shen Yang and has the ability to identify global optimal solution. It has two features over other algorithms: automatic subdivision and ability to deal with multimodality (Fister *et al*., 2013). FA has been successfully implemented to solve optimization problems such as traffic forecasting (Yusof *et al*., 2015) economic dispatch problem (Yang *et al*., 2012b). The operation of Firefly is based on two important factors; the light intensity and the attractiveness between fireflies. The light intensity of a firefly is related with the objective function f(x). The objective function can be formulated as maximization or minimization problem. On the other hand, the attractiveness, β, between fireflies is related with light intensity and it changes based on the distance between two fireflies as shown in step 7 where, in this study, β_0 is set to 1, Y which is the light absorption coefficient is set to 1 and r_{ij} represents the euclidean distance. The movement of firefly in step 6 is based on the position of

Step1: Generate Initial population of firefly randomly xi (i=1, 2,.., n),Light Intensity I at xi is determine by Objective function f(xi).
Step2: Define light absorption coefficient γ.
Step3: While (t < Max Generation)
Step4: For i=1 to N (N all fireflies)
Step5: For j=1 to N
Step6: If (Ii<Ij) { Xi = Xi + β(Xj - Xi) + α εi }*
Step7: β = β0 exp(-γrij2)
Step8: Evaluate new solutions and update light intensity.
Step9: End For i
Step10: End For j
Step11: Rank the fireflies and find the current global best g.*
Step12: End While

Fig. 1: The steps in Firefly algorithm

the less bright firefly X_i and the position of brightest firefly X_j while the random number α in the range (0, 1) where in this paper, it is set as 0.2. The steps in Firefly algorithm (Banati and Bajaj, 2013; Yang, 2010) are presented (Fig. 1).

In (Rui *et al*., 2012), the researchers proposed to investigate the ability of applying Firefly, Cuckoo, Bat and Wolf algorithms for clustering web intelligence data. In this study, we propose to integrate Firefly algorithm at each level tree construction in the Bisect algorithm. Such an approach is undertaken to solve the local optima problem.

MATERIALS AND METHODS

Proposed integrated Firefly algorithm with Bisect K-means (BISECT FA): In the proposed Bisect FA, at each level, the algorithm selects one cluster, C, that represents the whole dataset and classifies the objects into two partitions (C1 and C2) by choosing two best centers and assigning objects to the closest centers based on Firefly Algorithm (Mohammed *et al*., 2015) This process continues until it reaches the stopping condition which is a specific number of clusters. At each step of object classification, the cluster is evaluated using intra similarity criterion as shown in Eq. 1 where the cluster with maximum intra similarity is chosen as good cluster while the cluster that undergoes more classification in remaining levels has minimum intra similarity:

$$\text{Interasimilarity}\left(C_j\right) = \sum_{i=1}^{N_c} \left\| X_{i,j} * \text{Cenc}_j \right\|^2 \qquad (1)$$

In FA (Mohammed *et al*., 2015) initially, the number of firefly and number of clusters, k are specified. Each firefly will randomly choose two objects vector to be represented as initial centroids. Then, objects (i.e. documents) are assigned to the most similar (i.e., nearest) centroid. Evaluation of the clusters is later performed using objective function that is based on Average Distance between Documents and Center (ADDC) (He *et al*., 2006; Cui *et al*., 2005) as shown in Eq. 2:

$$F\left(X^*\right) = \min \sum\nolimits_{j=1}^{k} \frac{\frac{\sum\nolimits_{i=1}^{ni} ED\left(O_i, d_l\right)}{n_i}}{k} \qquad (2)$$

Where:
- k = Number of clusters
- n_i = Number of objects in cluster j
- ED = Euclidean distance between
- d_j = Documents in cluster j and
- O_j = Center of cluster j

The initial light of the firefly is based on ADDC objective function, where it equals one over ADDC value. Two fireflies will compete between each other based on their light brightness, where, the one with a brighter light will win, hence, forcing the less bright ones to move towards the winner. This process continues until it reaches a specific number of iteration. The winner (i.e., the brightness firefly) will carry information on the two best clusters. Evaluation on these clusters will be performed based on the intra similarity objective function using Eq. 1, cluster with the higher similarity is chosen for first level of Bisect FA while documents in the lower similarity cluster will be passed back to FA for another repetition of the clustering. The pseudo code of the proposed Bisect FA for text clustering.

RESULTS AND DISCUSSION

In order to evaluate the proposed Bisect FA algorithm for text clustering, experiments are conducted to compare the clustering result of proposed Bisect FA algorithm against the ones produced by Bisect K-means (Murugesan and Zhang, 2011a, b) K-means (Jain, 2010) and hybrid firefly algorithm with K-means (C-Firefly) (Rui *et al*., 2012). Experiments were undertaken in Matlab on windows 8 with a 2000 MHz processor and 4 GB memory. Each experiment was executed for 10 times and average values of the performance metrics are calculated.

The dataset utilized in this study is the one that has been widely utilized in information retrieval and text mining field which is the 20 newgroups (Bache and Lichman, 2013). The collection is obtained from UCI machine learning repository and is available at http://archive.ics.uci.edu/ml. The 20 newsgroups dataset contains 300 documents from 3 different classes- hardware, baseball and electronic, where each class includes 100 documents. The number of terms involve is 2275. Table 1 includes simple description of the data collection (Fig. 2).

Six performance metrics are used to evaluate the clustering result namely the ADDC (Murugesan and Zhang 2011; Cui *et al*., 2005). Purity, (Forsati *et al*. (2013); Murugesan and Zhang, 2011a), F-measure (Forsati *et al*., 2013; Murugesan and Zhang, 2011b), Entropy (Forsati *et al*., 2013; Murugesan and Zhang, 2011a, b), Davies-Bouldin Index DBI (Das *et al*., 2009) and Dunn Index DI (Das *et al*., 2009). A smaller value of ADDC, Entropy and DBI indicate good clustering while large values are required in purity, F-measure and DI.

Table 1: Description of data collection

Dataset	No. of documents	Total No. of classes	Min no. of documents in class	Max no. of documents in class	No. of Terms
20 Newsgroups	300	3	100	100	2275

Bisect FA:
Step1: Randomly choose two cluster centers.
Step2: Cluster using Firefly Algorithm FA.
Step3: Evaluate clusters using Eq.1.
Step4: If not reach number of clusters, choose the cluster that has smallest intra similarity for further classification
Step5: Repeat step1 until reach number of clusters.

Clustering using FA:
Step1: Generate initial population of firefly xi (i=1, 2, .., n),where each Firefly, randomly chooses 2 cluster centers.
Step2: For each Firefly do:
Step3: Assign each document to closest center.
Step4: Compute the Objective function f(x) which is based on Eq.2 of ADDC metric.
Step5: End For
Step6: Light Intensity, I, at xi is determined by f(xi).
Step7: Define light absorption coefficient, γ.
Step8: While (t < Max_iteration)
Step9: For i=1 to N (N is the number of fireflies)
Step10: For j=1 to N
Step11: If (Ii<Ij)
Step12: Move firefly i towards j using
{ Xi = Xi + β(Xj - Xi) + α εi }*
Step13: Calculate the attractiveness, β, using β = β0 exp(-γrij2)
Step14: Evaluate new solutions using Eq.2 and update light intensity.
Step15: Rank the fireflies and find the current global best solution (i.e the brightest firefly).

Fig. 2: The Pseudo code of proposed Bisect FA for text clustering

Table 2: Results of bisect FA vs. Bisect K-means vs. K-means vs. C-firefly

	Algorithms			
Metrics	Bisect FA	Bisect K-means	K-means	C-firefly
ADDC	**0.5878**	1.2602	0.6764	1.4436
Purity	0.34	0.3693	0.3463	**0.3737**
F-measure	**0.4992**	0.4871	0.4957	0.3743
Entropy	1.5744	**1.5616**	1.5746	1.5741
DBI	**0.9541**	6.02744	0.6685	14.0912
DI 0	9460	0.335	**3.8686**	0.1397

The best value is highlighted in 'bold '

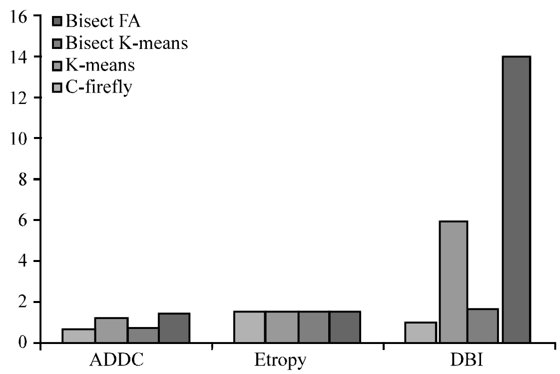

Fig. 3: A graphical representation of results (ADDC, Entropy and DBI): Bisect FA vs. Bisect K-means vs. K-means vs. C-firefly (lower value is the best)

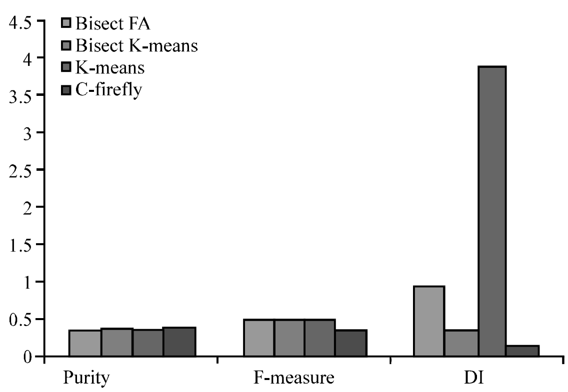

Fig. 4: A graphical representation of results (Purity, F-measure and DI): Bisect FA vs. Bisect K-means vs. K-means vs. C-firefly (higher value is the best)

Table 2 includes the results of the employed metrics for Bisect FA, Bisect K-means, K-means and C-Firefly. From data depicted in Table 2, it is learned that the proposed Bisect FA generates the best ADDC value which is 0.5878. It also obtains the best value for F-measure (0.4992) and DBI (0.9541) compared to the ones by Bisect K-means, K-means and C-firefly. Figure 3 and 4 illustrate the graphical representation of performance metrics among Bisect FA, Bisect K-means, K-means and C-firefly.

CONCLUSION

This study proposes a hierarchical text clustering algorithm based on integration between Bisect and firefly algorithm which is called Bisect FA. The aim of using Firefly algorithm is to perform a global search that later generates optimal clusters. In conducting the experiments, the performance of the proposed Bisect FA is analyzed on a benchmark dataset in text clustering which is the 20 newsgroups. Performance evaluation of the proposed Bisect FA is undertaken by comparing its results against Bisect K-means, K-means and C-firefly, using three different types of performance metrics, named as internal such as ADDC, external such as purity, F-measure and Entropy and relative metrics such as DBI and DI. The results indicate that Bisect FA is a better algorithm than Bisect K-means, K-means and C-firefly in terms of ADDC, F-measure and DBI. Hence, indicating that Bisect FA algorithm is a competitive method in hierarchical text clustering.

REFERENCES

Aliguliyev, R.M., 2009. Clustering of document collection: A weighting approach. Exp. Syst. Appli., 36: 7904-7916.

Bache, K. and M. Lichman, 2013. UCI machine learning repository. University of California, School of Information and Computer Science, Irvine, CA., USA.

Banati, H. and M. Bajaj, 2013. Performance analysis of firefly algorithm for data clustering. Int. J. Swarm Intell., 1: 19-35.

Beasley, D., D.R. Bull and R.R. Martin, 1993. An overview of genetic algorithms: Part 1, Fundamentals. Univ. Comput., 15: 58-69.

Bonabeau, E., M. Dorigo and G.X. Theraulaz, 1994. Swarm Intelligence: From Natural to Artificial Systems. Oxford University Press, New York.

Boussaid, I., J. Lepagnot and P. Siarry, 2013. A survey on optimization metaheuristics. Inform. Sci., 237: 82-117.

Cui, X., J. Gao and T.E. Potok, 2006. A flocking based algorithm for document clustering analysis. J. Syst. Archit., 52: 505-515.

Cui, X., T.E. Potok and P. Palathingal, 2005. Document clustering using particle swarm optimization. Proceedings of the IEEE Swarm Intelligence Symposium, Jun. 8-10, Pasadena, California, pp: 185-191.

Das, S., A. Abraham and A. Konar, 2009. Metaheuristic Clustering. Springer, Heidelberg, ISBN: 9783540921721, Pages: 254.

Dorigo, M., 1992. Optimization, learning and natural algorithms. Ph.D. Thesis, Politecnico di Milano, Italy.

Fister, I., X.S. Yang and J. Brest, 2013. A comprehensive review of firefly algorithms. Swarm Evol. Comput., 13: 34-46.

Fogel, D.B., 1994. Asymptotic convergence properties of genetic algorithms and evolutionary programming: Analysis and experiments. Cybern. Syst.: Int. J., 25: 389-407.

Forsati, R., M. Mahdavi, M. Shamsfard and M.R. Meybodi, 2013. Efficient stochastic algorithms for document clustering. Inform. Sci., 220: 269-291.

Glover, F., 1986. Future paths for integer programming and links to artificial intelligence. Comput. Operat. Res., 13: 533-549.

Han, J., M. Kamber and J. Pei, 2011. Data Mining: Concepts and Techniques. 3rd Edn., Morgan Kaufmann Publishers, USA., ISBN-13: 9780123814791, Pages: 744.

He, Y., S.C. Hui and Y. Sim, 2006. A Novel Ant-Based Clustering Approach for Document Clustering. In: Information Retrieval Technology, Ng, H.T., M.K. Leong, M.Y. Kan and D. Ji (Eds.). Springer, Berlin, Heidelberg, ISBN: 978-3-540-45780-0, pp: 537-544.

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern Recogn. Lett., 31: 651-666.

Kashef, R. and M.S. Kamel, 2009. Enhanced bisecting k-means clustering using intermediate cooperation. Pattern Recognition, 42: 2557-2569.

Kashef, R. and M.S. Kamel, 2010. Cooperative clustering. Pattern Recognition, 43: 2315-2329.

Kennedy, J. and R. Eberhart, 1995. Particle swarm optimization. Proceedings of the International Conference on Neural Networks, Volume 4, November 27-December 1, 1995, Perth, WA., USA., pp: 1942-1948.

Kirkpatrick, S., C.D. Gelatt Jr. and M.P. Vecchi, 1983. Optimization by simulated annealing. Science, 220: 671-680.

Mohammed, A.J., Y. Yusof and H. Husni, 2015. Basic firefly algorithm for document clustering. AIP Conf. Proc., Vol. 1691. 10.1063/1.4937068

Murugesan, K. and J. Zhang, 2011a. Hybrid bisect K-means clustering algorithm. Proceedings of the International Conference on Business Computing and Global Informatization, July 29-31, 2011, Shanghai, pp: 216-219.

Murugesan, K. and J. Zhang, 2011b. Hybrid hierarchical clustering: An experimental analysis. Technical Report No. CMIDA-HiPSCCS#001-11. University of Kentucky, USA.

Rokach, L. and O. Maimon, 2005. Clustering Methods. In: Data Mining and Knowledge Discovery Handbook, Maimon, O. and L. Rokach (Eds.). Springer, New York, pp: 321-352.

Rothlauf, F., 2011. Design of Modern Heuristics: Principles and Application. Springer-Verlag, Berlin, Heidelberg, ISBN: 9783540729624, Pages: 267.

Rui, T., S. Fong, X.S. Yang and S. Deb, 2012. Nature-inspired clustering algorithms for web intelligence data. Proceedings of the International Conferences on Web Intelligence and Intelligent Agent Technology, December 4-7, 2012, Macau, pp: 147-153.

Tang, R., S. Fong, X.S. Yang and S. Deb, 2012a. Integrating nature-inspired optimization algorithms to K-means clustering. Proceedings of the 7th International Conference on Digital Information Management, August 22-24, 2012, Macau, pp: 116-123.

Tang, R., S. Fong, X.S. Yang and S. Deb, 2012b. Wolf search algorithm with ephemeral memory. Proceedings of the 7th International Conference on Digital Information Management, August 22-24, 2012, Macau, pp: 165-172.

Yang, X.S. and S. Deb, 2009. Cuckoo search via Levy flights. Proceedings of the World Congress on Nature and Biologically Inspired Computing, December 9-11, 2009, Coimbatore, India, pp: 210-214.

Yang, X.S., 2010. A New Metaheuristic Bat-Inspired Algorithm. In: Nature Inspired Cooperative Strategies for Optimization, Gonzalez, J.R., D.A. Pelta, C. Cruz, G. Terrazas and N. Krasnogor (Eds.). Springer, Berlin, Germany, ISBN: 9783642125379, pp: 65-74.

Yusof, Y., F.K. Ahmad, S.S. Kamaruddin, M.H. Omar and A.J. Mohamed, 2015. Short term traffic forecasting based on hybrid of firefly algorithm and least squares support vector machine. Proceedings of 1st International Conference on Soft Computing in Data Science, September 2-3, 2015, Putrajaya, Malaysia, pp: 164-173.

Zaw, M.M. and E.E. Mon, 2013. Web document clustering using cuckoo search clustering algorithm based on levy flight. Int. J. Innov. Applied Stud., 4: 182-188.

Zhang, L., Q. Cao and J. Lee, 2013. A novel ant-based clustering algorithm using Renyi entropy. Applied Soft Comput., 13: 2643-2657.

Zhong, J., L. Liu and Z. Li, 2010. A Novel Clustering Algorithm based on Gravity and Cluster Merging. In: Advanced Data Mining and Applications, Cao, L., Y. Feng and J. Zhong (Eds.). Springer, Berlin, Heidelberg, ISBN: 978-3-642-17315-8, pp: 302-309.