Optimizing the Management of Knowledge Assets using Swarm Intelligence

Yuhanis Yusof[1], Fauziah Baharom[1] and Athraa Jasim Mohamed[2]

[1]*Universiti Utara Malaysia, Malaysia, {yuhanis@uum.edu.my, fauziah@uum.edu.my}*
[2]*University of Technology, Baghdad, Iraq, {10872@uotechnology.edu.iq}*

ABSTRACT

Knowledge assets are the knowledge drivers of an organization's success and they can be of structured, unstructured, tacit or explicit knowledge. Explicit knowledge are realized as documents and these documents need to be tagged in order to ease the process of retrieval. Such document grouping process can help an organisation to meet legal and regulatory requirements for retrieving specific information in a set timeframe. Current document clustering techniques rely on a pre-defined value of *k* (number of clusters). Hence, the produced clusters will be of different quality. This study presents the employment of swarm intelligence algorithm, i.e Firefly Algorithm, to automatically cluster text document without the use of *k* value. Experimental results shows that the performance of the algorithm is better compared to the benchmark methods. The number of obtained clusters are the same as the ones defined in the data collection while the purity value for three out of four datasets are higher than the benchmark methods. Hence, this indicates that the proposed swarm intelligence based clustering facilitates the grouping of knowledge assets. By having an automated document clustering, tagging the document with their appropriate label will help organization to better manage their knowledge assets.

Keywords: Knowledge assets, document clustering, swarm intelligence, firefly algorithm, data mining.

I INTRODUCTION

Knowledge assets are the knowledge drivers of an organization's success (Eppler, 2008). Within business and knowledge management, two types of knowledge are usually defined, namely explicit and tacit knowledge. The former refers to codified knowledge, such as that found in documents, while the latter refers to non codified and often personal/experience-based knowledge. Both of these two types are knowledge assets as they carry certain value in the business operation of the organization. The more structured a knowledge asset is, the easier it is to manage (update, share, remove, etc.) that knowledge, both internally and externally. Although the growing significance of intangible assets was recognized during the second half of the 20th century, it was not until the last two decades of the 20th century that concepts of knowledge management and organizational learning became popular (Masic, Nesic, Nikolic, & Dzeletovic, 2017).

Document clustering organizes text documents as clusters; similar documents are in one group and dissimilar ones in another group (Xinwu, 2010). Document clustering is applicable in document organization and browsing which the hierarchical approach can be very beneficial for documents to be browsed systematically. In addition, it discovers a hidden pattern based on the similarities between the documents (Aggarwal & Zhai, 2012).

Various methods have been reported to contribute in document clustering and this includes the employment of swarm intelligence. Swarm Intelligence is an emerging field in the optimization research community. It is a subset of Evolutionary Computing which is motivated from the intelligent behavior of insects or animal (Karaboga, et al., 2012). The term swarm implies the aggregation of insects or animals such as fish schools, bird flocks and others. Since decades ago, many swarm algorithms have been presented and this includes Firefly Algorithm (FA) which was developed by Xin-She Yang in 2007. Firefly algorithm has been used in many applications, such as economic emission load dispatch problem (Apostolopoulos & Vlachos, 2011; Yang, Hosseini, & Gandomi, 2012), speech recognition (Hassanzadeh, Faez, & Seyfi, 2012), image segmentation (Hassanzadeh, Vojodi, & Moghadam, 2011), reliability-redundancy allocation problem (Dos Santos Coelho, de Andrade Bernert, & Mariani, 2011), semantic web service composition (Pop et al., 2011), data classification (Nandy, Sarkar, & Das, 2012), anomaly detection (Adaniya, Abr˜ao, & Proenc¸a Jr., 2013), and parallel and distributed systems (Falcon, Almeida, & Nayak, 2011).

This study presents the use of FA to auomatically cluster knowledge asset (i.e text document) into hierarchical clustering. The determination of

optimal number of clusters (i.e k value) and the creation of the clusters is performed by the Weight-based Firefly Algorithm with relocating and merging algorithm (WFARM) which was detailed by Mohammed (2016).

II RELATED WORK

The literature for this study is built upon three major components; knowledge assets, document clustering and firefly algorithm. The required discussion is presented in the following subsections.

A. Knowledge Assets

According to Miller and Shamsie (1996), knowledge has long been recognized as a valuable resource for organizational growth and sustained competitive advantage, especially for organizations competing in an uncertain environment. Nonaka, Toyama & Konno (2000) defined knowledge assets as oraganization-specific resources that are indispensable to creating value for the organization. They introduce four types of knowledge assets; experiential knowledge assets, conceptual knowledge assets, systemic knowledge assets and routine knowledge assets (Figure 1).

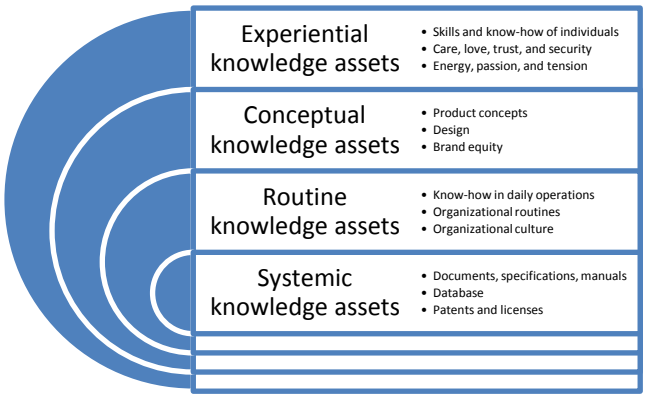

Figure 1: Categories of Knowledge Assets (Adopted from Nonaka, Toyama and Konno (2000))

The study by Freeze and Kulkarni (2007) presents the components of knowledge capabilities that includes knowledge documents (Figure 2).

Attention to knowledge asset classification initially started in early 2000 as various study focuses to identify step and procedures to classify knowledge assets. The study by Epler (2005) proposes the evaluation method, knowledge map and education program to facilitate knowledge asset classification.

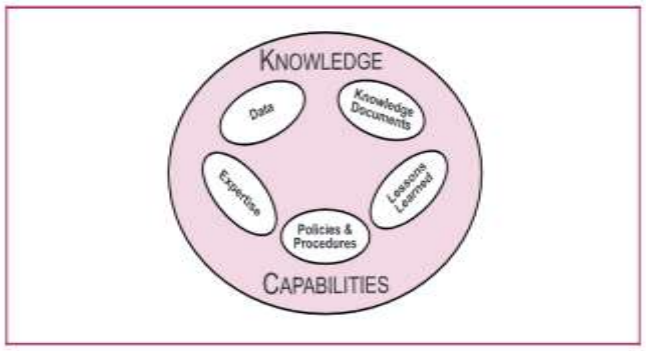

Figure 2. Knowledge Management Framework (adopted from Freeze and Kulkarni (2007))

B. Document Clustering

To date, there exist various methods on text clustering. These include the development of enhanced algorithms and hybridization of existing clustering algorithms. In general, the methods can be divided into five types: Partitional clustering, Hierarchical clustering, Density-based clustering, Model-based clustering and Grid-based clustering (Zhang, Cao, & Lee, 2013). Regardless of the types, the clustering process will either be of static or dynamic. The first refers to the use of k value (pre-defined number of clusters) while the latter does not require the information.

Partitional clustering algorithms divide a dataset into groups based on the inter-similarity between documents. The most popular and efficient partitional clustering algorithm is K-means (Jain, 2010). The implementation of K-means generates problems that include the randomly selected initial centroids. Existing studies (Gu, Zhou, & Chen, 2009; Mishra, Nayak, Rath, & Swain, 2012; Yao, Pi & Cong, 2012) indicate that different initial centroids produce different quality of clusters. Hence, this indicates that it is important to accurately determine the initial centroids.

Density-based clustering algorithm is a technique to construct clusters based on the dense regions of objects in high-dimensional space that are isolated by low density areas. Density-Based Spatial Clustering of Application with Noise (DBSCAN) is the most known density-based method used for data clustering (Ester, Kriegel, Sander, & Xu, 1996). It randomly selects points and finds the neighborhood by using query. A cluster is constructed based on these points and each neighbor examined to see if it can be included in the cluster (Chehreghani, Abolhassani, & Chehreghani, 2008).

The grid-based clustering approach uses a multi-resolution grid data structure. It divides data space into several levels of cells. The process of clustering is performed inside these cells. The parameters of higher level cells can be computed using lower level

cells. The quality of clustering is based on the number of cells in lower level cells. If it is too coarse, this will lead to the quality of cluster being reduced (Han & Kamber, 2011). In addition, grid clustering does not have a relationship between neighbors, which means there are no children and parent cells to be represented hierarchically. There are various recognized grid-based approaches such as STING (Wang, Yang, & Muntz, 1997) and OptiGrid (Hinneburg & Keim, 1999).

Model-based clustering is the method that tries to optimize the fit between some mathematical models and data. The Self Organizing Map (SOM) is a type of model-based clustering and Neural Network which was presented by Kohonen (1998). The SOM algorithm has been used in many applications, like semantic map, clustering, and so on (Yin, Kaku, Tang, & Zhu, 2011). Model-based clustering is sensitive to the initial selection of weight vector, as well as to its different parameters, such as the learning rate and neighborhood radius (Rokach & Maimon, 2005).

C. Firefly Algorithm

Fireflies are winged beetles which produce short and rhythmic flashes. The flashing light is generated by the bioluminescence process. Firefly uses bioluminescence to attract mates or prey. The Firefly Algorithm (FA) was developed by Xin-She Yang in 2007 at Cambridge University. FA has two important issues: light intensity and attractiveness. For maximum optimization problems, the light intensity I of a firefly at a particular location x, termed as I(x), can be determined by objective function f(x). The attractiveness β is relative. It changes depending on the distance between two fireflies (Yang & He, 2013; Yang, 2010a, 2010b).

Firefly Algorithm has been implemented in many optimization problems in different topics, such as speech recognition (Hassanzadeh, Faez, & Seyfi, 2012), image segmentation (Hassanzadeh, Vojodi, & Moghadam, 2011; Horng & Jiang, 2010), reliability-redundancy allocation problems (dos Santos Coelho, de Andrade Bernert, & Mariani, 2011), discrete optimization problems (Sayadi, Hafezalkotob, & Naini, 2013), semantic web service composition (Pop et al., 2011), data classification (Nandy, Sarkar, & Das, 2012), anomaly detection (Adaniya Abr˜ao & Proenc¸a Jr., 2013), parallel and distributed systems (Falcon, Almeida, & Nayak, 2011), mobile network (Bojic, Podobnik, Ljubi, Jezic, & Kusek, 2012), and economic dispatch problems (Yang, Hosseini, & Gandomi, 2012). In all of the previous fields, Firefly Algorithm has successfully identified the optimal solution.

III METHODOLOGY

This study is implemented by performing four (4) main phases; data collection, data representation, document clustering and clustering evaluation. In phase one, secondary data is utilized as they are commonly used in clustering experiments. In specific, the TREC collection that includes TR11, TR12, TR23 and TR45 has been obtained from CLUTO toolkit (Karypis, 2002). TR11 includes 414 documents from nine different classes and the number of terms is 6429. TR12 contains 313 documents from eight different classes with 5804 terms. TR23 includes 204 documents distributed in six classes and the number of terms is 5832. The last dataset, TR45, contains 690 documents from ten classes and consists of 8261 terms.

In order to represent the collected data, the Vector Space Model (VSM) is utilized. In VSM, let $D = \{D1, D2, \ldots, D_n\}$ as the document collection and n represents the number of documents in the collection. Let $T = \{T1, T2, \ldots, T_m\}$ be the terms in each documents and m represents the number of terms. In vector space model, the document *D* is represented as a vector in the m dimensional space (Aliguliyev, 2009a, 2009b). The vector D is related with the terms by a degree value. In this study, the TFIDF is employed in the VSM to indicate the importance of words that exist in each document as well as in the whole collection.

The document clustering phase includes the employment of WFA_{RM} (Mohammed, 2016) on the TREC collection. Experiments were performed in Matlab simulation platform and the obtained results were compared against the one produced by exisiting clustering algorithms (i.e K-Means, FAK-means and BatK-means).

IV RESULTS

In order to evaluate the effectivess of WFA_{RM} (Mohammed, 2016) , discussion of the results is based on 2 metrics; number of clusters and purity of clusters. Table 1 represents the number of clusters obtained by WFA_{RM}, K-means, FAK-mean and BatK-means.

Table 1. ***Numbers Of Clusters***

Datasets	Number of clusters of algorithms			
	WFA_{RM}	K-means	FAK-means	BatK-means
TR11 (414 documents and 9 classes)	9	9	9	9
TR12 (313 documents and 8 classes)	8	8	8	8
TR23	6	6	6	6

(204 documents and 6 classes)				
TR45 (690 documents and 10 classes)	10	10	10	10

Based on the data, it noted that all clustering algorithms have obtained the same number of clusters and they as of required (i.e same with the data collection). This shows that all of the methods are capable to group the documents. Hence there is a need to compare the quality of the obtained clusters. Data in Table 2 includes the purity value (i.e average and standard deviation) of the clusters. The best value is highlighted as bold. To compute purity , each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned documents and dividing by *N*.

Table 2. Purity Value (Average And Standard Deviation(SD))

Datasets	Algorithms	Average	SD
TR11 (414 documents and 9 classes)	WFA_{RM}	**0.6810**	0.0224
	K-means	0.3860	0.0722
	FAK-means	0.3239	**0.0060**
	BatK-means	0.6657	0.0793
TR12 (313 documents and 8 classes)	WFA_{RM}	0.4752	**0.0022**
	K-means	0.4370	0.0827
	FAK-means	0.3013	0.0065
	BatK-means	**0.5808**	0.0639
TR23 (204 documents and 6 classes)	WFA_{RM}	**0.6266**	**0.0032**
	K-means	0.5451	0.0559
	FAK-means	0.4495	0.0072
	BatK-means	0.5956	0.0524
TR45 (690 documents and 10 classes)	WFA_{RM}	**0.6355**	0.0161
	K-means	0.4416	0.0988
	FAK-means	0.2422	**0.0056**
	BatK-means	0.6083	0.0789

As shown in Table 2, the WFA_{RM} algorithm generates the highest average purity in most datasets (TR11, TR23 and TR45) compared to other algorithms. The BatK-means generated the highest average purity of 0.5808 in the TR12 dataset. In addition, the standard deviation of the WFA_{RM} algorithm is smaller than other methods in half of the datasets.

The study further evaluate the clustering performance by performing T-test to undertstand whether there is a significant difference of the obtained metrics. The test is to determine which hypothesis will be accepted:

H0: Mean of (WFA_{RM}) = Mean of (benchmark method)
H1: Mean of (WFA_{RM}) ≠ Mean of (benchmark method)

In Table 3, the associated P-value (sig 2-tailed test) using average Purity is illustrated. Since the P-value between WFA_{RM} and any other methods is smaller than (0.05), the null hypothesis is rejected, the mean of any metrics for WFA_{RM} and any benchmark methods is the same while the alternative hypothesis is accepted to conclude that there is a significant difference in the mean of purity metric for WFA_{RM} and any benchmark methods. This excludes the P-value between WFA_{RM} and BatK-means (bold value in Table 3) in two datasets; TR11 and TR45. which are (0.3136 and 0.3166, 0.0697 and 0.0740).

Table 3: T-test Result

Datasets	Algorithms	P-value using average Purity (sig 2 tailed)	
		Equal variances assumed	Equal variances not assumed
TR11	WFA_{RM} : K-means	3.5848E-29	1.7263E-21
	WFA_{RM} : FAK-means	2.4134E-62	2.8869E-40
	WFA_{RM} : BatK-means	**0.3136**	**0.3166**
TR12	WFA_{RM} : K-means	0.0141	0.0171
	WFA_{RM} : FAK-means	5.0969E-75	1.7287E-50
	WFA_{RM} : BatK-means	1.0808E-12	5.8565E-10
TR23	WFA_{RM} : K-means	6.5395E-11	7.984E-09
	WFA_{RM} : FAK-means	6.2078E-72	3.5164E-53
	WFA_{RM} : BatK-means	0.00199	0.00299
TR45	WFA_{RM} : K-means	3.3588E-15	9.2252E-12
	WFA_{RM} : FAK-means	1.7519E-72	4.6537E-49
	WFA_{RM} : BatK-means	**0.0697**	**0.0740**

V CONCLUSION

Knowledge asset in particular the ones realized as explicit knowledge need to be properly organized in order to ease information retrieval. As organization relies on knowledge asset to make important decisions and strategic planning, the automation of document clustering is important in an organization. In this study, the employment of one of the swarm intelligence algorithm, that is the Firefly Algorithm, is presented. It is learned that the automation of k-value determination performed in WFA_{RM} has not only removed human involvement but also produced quality clusters. Such an achievement indicates the possibility of uplifting knowledge management to a higher level by incorporating swarm intelligence that has the capability to optimize knowledge assets.

ACKNOWLEDGMENT

This study has been supported by Exploratory Research Grant Scheme (s/o 12827).

REFERENCES

Adaniya, M. H. A. C., Abr˜ao, T., & Proenc¸a Jr., M. L. (2013). Anomaly Detection Using Metaheuristic Firefly Harmonic Clustering. *Journal of Networks*, *8*(1), 82–91. Retrieved from doi:10.4304/jnw.8.1.82-91

Aliguliyev, R. M. (2009a). Clustering of document collection-A weighted approach. *Elsevier, Expert Systems with Applications*, *36*(4), 7904–7916. Retrieved from doi: 10.1016/j.eswa.2008.11.017

Aliguliyev, R. M. (2009b). Performance evaluation of density-based clustering methods. *Elsevier, Information Sciences*, *179*(20), 3583–3602. Retrieved from doi: 10.1016/j.ins.2009.06.012

Aggarwal, C. C., & Zhai, C. X. (2012). A survey of text clustering algorithms. In *In Mining Text Data, Springer US* (pp. 77–128). Retrieved from doi:10.1007/978-1-4614-3223-4_4

Apostolopoulos, T., & Vlachos, A. (2011). Application of the Firefly Algorithm for Solving the Economic Emissions Load Dispatch Problem. *International Journal of Combinatorics*, *Volume 201*, 23 pages. Retrieved from doi:10.1155/2011/523806

Chehreghani, M. H., Abolhassani, H., & Chehreghani, M. H. (2008). Improving density-based methods for hierarchical clustering of web pages. *Elsevier, Data & Knowledge Engineering*, *67*(1), 30–50. Retrieved from doi: 10.1016/j.datak.2008.06.006

Dos Santos Coelho, L., de Andrade Bernert, D. L., & Mariani, V. C. (2011). A chaotic firefly algorithm applied to reliability-redundancy optimization. In *2011 IEEE Congress on Evolutionary Computation (CEC)* (pp. 517–521). New Orleans, LA. Retrieved from doi:10.1109/CEC.2011.5949662

Eppler, M. J. (2008). A Process-Based Classification of Knowledge Maps and Application Examples. *Knowledge and Process Management*, *15*(1), 59-71.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press.* (pp. 226–231).

Falcon, R., Almeida, M., & Nayak, A. (2011). Fault Identification with Binary Adaptive Fireflies in Parallel and Distributed Systems. In *2011 IEEE Congress on Evolutionary Computation (CEC)*, (pp. 1359–1366). New Orleans, LA: IEEE Explore. Retrieved from doi:10.1109/CEC.2011.5949774

Freeze, R. D., & Kulkarni, U. R. (2007). Knowledge Management Capability: Defining Knowledge Assets. Journal of Knowledge Management, 11(6), 94-109.

Gu, J., Zhou, J., & Chen, X. (2009). An Enhancement of K-means Clustering Algorithm. In *IEEE, International Conference on Business Intelligence and Financial Engineering* (pp. 237–240). Beijing: IEEE. Retrieved from doi: 10.1109/BIFE.2009.204

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques, 3rd edition*. *The Morgan Kaufmann Series in Data Management Systems* (p. 744 pages). Morgan Kaufmann.

Hassanzadeh, T., Faez, K., & Seyfi, G. (2012). A Speech Recognition System Based on Structure Equivalent Fuzzy Neural Network Trained by Firefly Algorithm. In *International Conference on Biomedical Engineering (ICoBE)* (pp. 63–67). Penang: IEEE Explore. Retrieved from doi:10.1109/ICoBE.2012.6178956

Hassanzadeh, T., Vojodi, H., & Moghadam, A. M. E. (2011). An Image Segmentation Approach Based on Maximum Variance Intra-Cluster Method and Firefly Algorithm. In *Seventh International Conference on Natural Computation (ICNC)* (Vol. 3, pp. 1817–1821). Shanghai: IEEE Explore. Retrieved from doi:10.1109/ICNC.2011.6022379

Hinneburg, A., & Keim, D. (1999). Optimal Grid-Clustering: Towards Breaking the Curse of Dimensionality in High-Dimensional Clustering. In *Proceedings of the 25th International Conference on Very Large Data Bases* (pp. 506–517). Morgan Kaufmann Publishers Inc.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Elsevier, Pattern Recognition Letters*, *31*(8), 651–666. Retrieved from doi: 10.1016/j.patrec.2009.09.011

Kohonen, T. (1998). The self-organizing map. *Elsevier, Neurocomputing*, *21*(1-3), 1–6. Retrieved from doi: 10.1016/S0925-2312(98)00030-7

Masic, B., Nesic, S., Nikolic, D., & Dzeletovic, M. (2017). Evolution of Knowledge Management. *Industrija*, *45*(2), 127-147.

Miller, D. and Shamsie, J. (1996), The resource-based view of the firm in two environments: The Hollywood Film Studios from 1936 to 1965, Academy of Management Journal, 39(3), 519-543.

Mishra, B. K., Nayak, N. R., Rath, A., & Swain, S. (2012). Far Efficient K-Means Clustering Algorithm. In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics* (pp. 106–110). ACM. Retrieved from doi:10.1145/2345396.2345414

Mohammed, A. J. (2016). Adaptive Firefly Clustering Algorithm for Hierarchical Text Clustering. (PhD), Universiti Utara Malaysia.

Nandy, S., Sarkar, P. P., & Das, A. (2012). Analysis of a Nature Inspired Firefly Algorithm based Back-propagation Neural Network Training. *International Journal of Computer Applications*, *43*(22), 8–16. Retrieved from doi:10.5120/6401-8339

Rokach, L., & Maimon, O. (2005). *Clustering Methods, Data Mining and Knowledge Discovery Handbook*. *Springer* (pp. 321–352.).

Pop, C. B., Chifu, V. R., Salomie, I., Baico, R. B., Dinsoreanu, M., & Copil, G. (2011). A Hybrid Firefly-inspired Approach for Optimal Semantic Web Service Composition. *Scientific International Journal for Parallel and Distributed Computing*, *12*(3), 363–369. Retrieved from retrived from: http://www.scpe.org/index.php/scpe/article/view/730/0

Xinwu, L. (2010). Research on Text Clustering Algorithm Based on Improved K-means. In *International Conference On Computer Design And Appliations (ICCDA 2010)* (Vol. 4, pp. V4–573 – V4–576). Qinhuangdao: IEEE. Retrieved from doi: 10.1109/ICCDA.2010.5540727

Wang, W., Yang, J., & Muntz, R. (1997). STING : A Statistical Information Grid Approach to Spatial Data Mining. In *VLDB '97 Proceedings of the 23rd International Conference on Very Large Data Bases* (pp. 186–195). Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.

Yang, X. S. (2009). Firefly Algorithms for Multimodal Optimization. In O. Watanabe & T. Zeugmann (Eds.), *Stochastic Algorithms: Foundations and Applications* (pp. 169–178). Springer Berlin Heidelberg. doi:10.1007/978-3-642-04944-6_14

Yang, X. S. (2010a). Firefly Algorithm, Stochastic Test Functions and Design Optimisation. *Int. J. Bio-Inspired Computation*, *2*(2), 78–84.

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

Yang, X. S., Hosseini, S. S. S., & Gandomi, A. H. (2012). Firefly Algorithm for solving non-convex economic dispatch problems with valve loading effect. *Elsevier, Applied Soft Computing*, *12*(3), 1180–1186. Retrieved from doi:10.1016/j.asoc.2011.09.017

Yao, M., Pi, D., & Cong, X. (2012). Chinese text clustering algorithm based k-means. In *2012 International Conference on Medical Physics and Biomedical Engineering (ICMPBE2012)* (Vol. 33, pp. 301–307). Elsevier. Retrieved from doi: 10.1016/j.phpro.2012.05.066, Available online at www.sciencedirect.com

Yin, Y., Kaku, I., Tang, J., & Zhu, J. (2011). *Data Mining Concepts, Methods and Application in Management and Engineering Design*. Springer-Verlag London.

Zhang, L., Cao, Q., & Lee, J. (2013). A novel ant-based clustering algorithm using Renyi entropy. *Elsevier, Applied Soft Computing*, *13*(5), 2643–2657. Retrieved from doi:10.1016/j.asoc.2012.11.022