International Journal of Information & Computation Technology.
ISSN 0974-2239 Volume 4, Number 17 (2014), pp. 1847-1860

http://www. irphouse.com

A Review ON K-means DATA Clustering APPROACH

Shraddha Shukla and Naganna S.

School of Computing Sciences and Engineering,
Galgotias University, Greater Noida, India

Abstract

In data mining, clustering is a technique in which the set of objects are assigned to a group called clusters. Clustering is the most essential part of data mining. K-means clustering is the basic clustering technique and is most widely used algorithm. It is also known as nearest neighbor searching. It simply clusters the datasets into given number of clusters. Numerous efforts have been made to improve the performance of the K-means clustering algorithm. In this paper we have been briefed in the form of a review the work carried out by the different researchers using K-means clustering. We have discussed the limitations and applications of the K-means clustering algorithm as well. This paper presents a current review about the K means clustering algorithm.

Keywords: K-means clustering, nearest neighbor searching, clusters and data mining.

I. INTRODUCTION

Due to the increased availability of computer hardware and software and the rapid computerization of business, large amount of data has been collected and stored in databases. Researchers have estimated that amount of information in the world doubles for every 20 months.

However raw data cannot be used directly. Its real value is predicted by extracting information useful for decision support. In most areas, data analysis was traditionally a manual process. When the size of data manipulation and exploration goes beyond human capabilities, people look for computing technologies to automate the process [12].

Data mining is one of the youngest research activities in the field of computing science and is defined as extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data.

Data mining is applied to gain some useful information out of bulk data. There are number of tools and techniques provided by researchers in data mining to obtain the pattern out of data. Different patterns can be mined by classification, clustering, association rules, regression, outlier analysis, etc. [77].

II. K-MEANS CLUSTERING

K-means clustering is most widely used clustering algorithm which is used in many areas such as information retrieval, computer vision and pattern recognition. K-means clustering assigns *n* data points into *k* clusters so that similar data points can be grouped together. It is an iterative method which assigns each point to the cluster whose centroid is the nearest. Then it again calculates the centroid of these groups by taking its average. The algorithm 1 shows the basic approach of K-means clustering [14].

1: An initial clustering is created by choosing k random centroids from the dataset.
2: For each data point, calculate the distance from all centroids, and assign its membership to the nearest centroid.
3: Recalculate the new cluster centroids by the average of all data points that are assigned to the clusters.
4: Repeat step 2 until convergence.

Algorithm 1 K-Means Clustering [14]

The working of Algorithm 1 can be explained clearly with the help of an example, which is shown on Figure 2.

Figure 2 shows the graphical representation for working of K-means algorithm. In the first step there are two sets of objects. Then the centroids of both sets are determined. According to the centroid again the clusters are formed which gave the different clusters of dataset. This process repeats until the best clusters are achieved.

There are abundant tools available for data mining. Some of them are Rapid Miner, R, Knime, Own Code, Weka Or Pentaho, Statistica, Sas Or Sas Enterprise Miner, Orange, Tanagra, And Matlab.

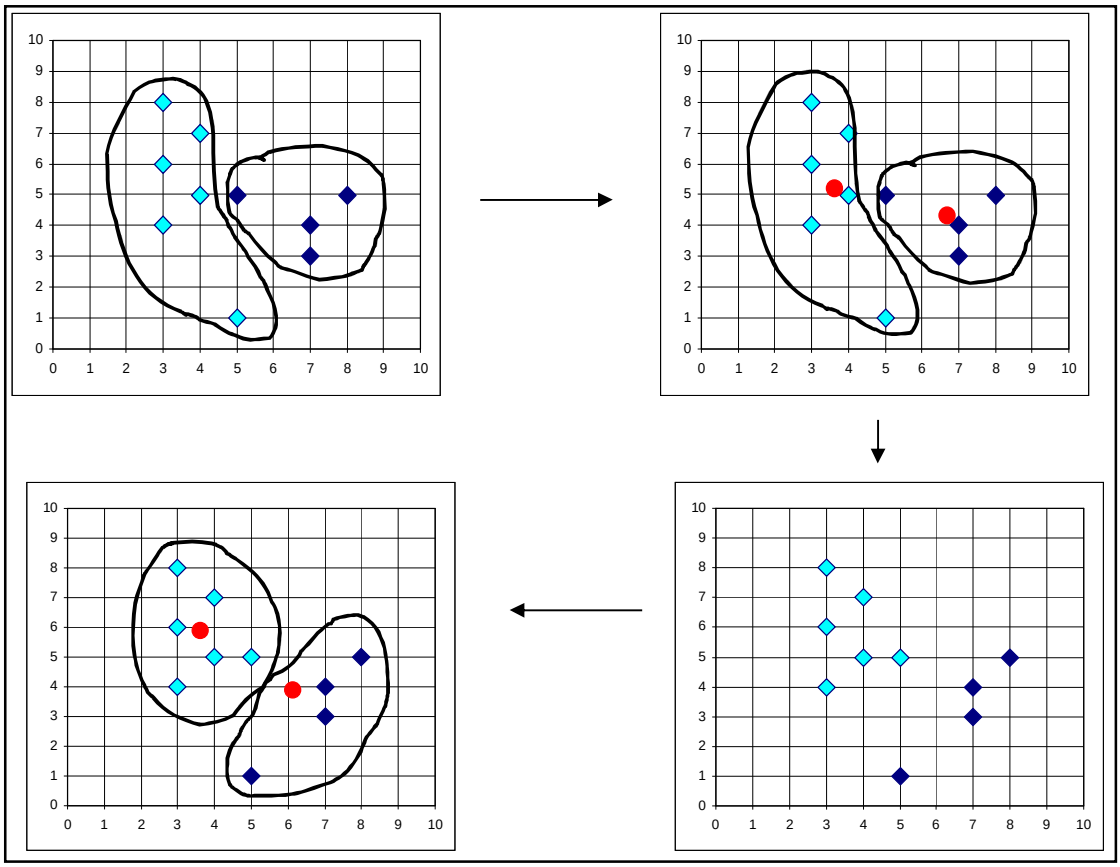

Figure 2 Working of K-means clustering algorithm

III. EVOLUTION

K-means was introduced by James MacQueen in 1967 [26]. It is observed that a lot of work has been done in this field. In the time frame of 1967 to 1998, all the research work was related to the introduction of K-means in clustering area. After this all the modifications and improvements were started on K-means clustering. Alsabti et al [27] has given an efficient clustering way by making a pattern in a k-d tree so that one can find the desired pattern easily. Kamal et al [33] has introduced an algorithm which uses labeled and unlabeled documents based on expectation machine and a classifier. They have concluded that their algorithm shows improvements in classification results. Mike et al [40] have presented the implementation of K-means clustering on an Annapolis wildstar that exhibited a speed up of two orders of magnitude.

Kiri et al [36] have developed a general method for having background knowledge out of K-means clustering algorithm in the form of constraints. Tapas et al [62] have presented an implementation of Lloyd's K-means clustering algorithm, which they termed as filtering algorithm. Cheung [75] has given a generalized way of K-means clustering, originally was given by [26].by this algorithm, correct clustered can be formed without initially known number of clusters. A simple study of the influence of

the semantics in the IR using LSI and K-means clustering technique has been done by Jimenez et al [7].

Modha and Spangler [10] have obtained a structure for integrating multiple, diverse feature spaces in K-means clustering algorithm. Tian et al [61] gives a study on parallel K-means clustering algorithm and presented a superior initial centers method to reduce the number of actions required while grouping. Pham et al [11] has worked on the number of k used in K-means clustering. They have concluded different number of clusters for different datasets. A survey was conducted by Xindong wu et al [71]. In their survey they have given the top 10 algorithms in data mining with their limitations, current and future works.

Xiong et al [18] has provided the results of the effect of skewed data distribution on K-means clustering. They have given an organized study of K-means and cluster validation measures from a data distribution perspective. In fact, their focus was on characterization of the relationships between data distribution and K-means clustering in addition to entropy measure and f-measure.Xiuyun Li et al [68] have proposed an improved K-means clustering which uses fuzzy feature selection. They used feature important factor to get the contribution of all the features in clustering. Osama Abu Abbas [49] has given a comparison between some of the data clustering algorithms.

Yan Zhu et al [79] has proposed a new method in which clustering initialization has been done using clustering exemplars produced by affinity propagation. They have also minimized the total squared error g the clusters. Taoying Li et al [63] has given a new approach towards fuzzy K-means clustering and has concluded with higher efficiency, greater precision and reduced amount of calculation.Viet-vu vu et al [65] has proposed an efficient algorithm for active seed selection which is based on min max approach that favors the coverage of whole dataset. Zhu and wang [25] has given an improved clustering algorithm with the use of genetic algorithm.

Oyelade et al [50] has implemented K-means clustering algorithm to analyze the academic performances of students on the basis of some pre set measures. Wu and Yao [70] have applied their improved algorithm on clustering analysis of transit data with same site name and different locations.

Yufen Sun et al [76] has presented a general K-means clustering to identify natural clusters in datasets. They have also shown high accuracy in their results. Wang and Yin [66] have shown that their algorithm has overcome the deficiencies of original K-means clustering and has higher accuracy. Li Xinwu [38] has given higher accuracy and better stability by improving the clustering algorithm. Napoleon and Lakshmi [8] has analyzed the time taken for execution by the original K-means clustering and their proposed K-means algorithm.

Hesam et al [17] has given improvements for guided K-means algorithm so that astrophysics data bases can be handled. Shi Na et al [54] present a simple way to assign data points to clusters. Their improved algorithm works in O (nk) time with high accuracy. Shamir and Tishby [51] have concluded that K-means does not break down in the large sample regime. Mark and Boris [42] address the most controversial issue of clustering i.e. the selection of right number of clusters.

In Honda et al [29], they have proposed a method for PCA guided K-means clustering of incomplete datasets. They have concluded that a PCA guided K-means

clustering is more robust than K-means clustering with PDS. Sathya et al [43] has given an approach to efficiently retrieve the search of clusters, in which comparison is based on the similarity of documents and co occurrence term, of the query. Xueyi Wang [69] has proposed a new algorithm kMkNN for the nearest neighbor searching problem. He has considered implementation of K-means clustering and triangle inequality. Ren and Fan [55] has introduced a K-means clustering approach based on coefficient of variation. They show how their approach can generate better results than K-means clustering algorithm.

Son and Anh [48] have compared two different methods for center initialization. One is kd tree and another is CF tree. Thomas A. Runkler [58] has focused on partially supervised clustering and introduced a partially supervised k – harmonic means clustering. Fatta et al [15] has proposed a decentralized algorithm for K-means. They concluded that their proposed method is practical and accurate. Murugesan and Zhang [32] introduced hybrid algorithm for K-means clustering. They uses two approaches top down and bottom up as bisect K-means and UPGMA respectively. Sarma et al [60] have proposed a fast method for K-means clustering which was useful for large datasets. They have come to know that their approach speed up the kernel K-means clustering. Wang and Su [23] have modified the clustering algorithm and showed their test results on iris, wine and abalone datasets. They have improved the algorithm with respect to the time and accuracy of the results.

Tripathy et al [6] have proposed a method for traditional kernel based K-means clustering algorithm which will later use the rough set concept for updation in the centroid value. Abhay et al [1] gave a model for predicting the outcome as yes or no in K-means clustering on weather data. The thought of maximum triangle rule was proposed by Feng et al [21] to optimize K-means clustering algorithm. They have overcome the shortcoming of K-means by introducing KMTR for the improvement in clusters. Ekasit et al. [14] have proposed parallel K-means on GPU clusters. They use the task pool model for dynamic load balancing to distribute workload equally on different GPUs installed in the clusters so as to improve the performance of the parallel K-Means at the inter-node level.

Jing et al. [22] presented a simple, easily parallelized and efficient K-means clustering algorithm via closure. Zhang et al [78] present a clustering algorithm on self adaptive weights. Their experimental result shows that it is more accurate and stable. Mahmud et al [45] have shown improvement in algorithm by taking weighted average to overcome the initial seed point limitation. They have also reduced the number of iterations for the clustering procedure. Wang et al [67] has improved the K-means clustering using the density concept. They succeeded in obtaining increased clustering precision and criterion function E.

Lee and Lin [56] have designed a selection and erasure K-means algorithm. The authors achieved increase in the efficiency with large no. of clusters. Patil and Vaidya [77] carried out a review on different clustering techniques. Cheng et al [73] proposed a system named cluchunk. This system clusters the unlabeled web data which incorporate chunklet information. Bikram et al. [5] have made some improvements in the traditional K-means clustering algorithm and used DI and DBI parameters for clustering validation. Ellis et al [13] has implemented a quantum based K-means

clustering, which shows the improvements in accuracy and precision. Nuno et al. [47] have improved the text clustering approach via K-means using overlapping community structures of a network of tags.

Anoop and Satyam [2] made a survey of recent clustering techniques in data mining. Vijayalakshmi and Renuka [46] discussed different methodologies and parameters associated with different clustering algorithms. They also discussed on issues in different clustering algorithms used in large datasets. Kurt et al [30] has presented spherical K-means clustering and suitable extensions. They also introduced R extension package skmeams. Deepti et al. [9] made a study on different clustering algorithms with introduction, application, limitations, and their requirements. Maryam et al. [39] have made an analysis on all clustering algorithms to choose the best algorithm for identifying duplicate entities. Rupali and Suresh [52] proposed an improved K-means clustering algorithm for two dimensional data.

Biggio et al. [4] has presented an approach to evaluate clustering algorithm's security in adversial settings. Khadem et al. [12] provided a survey on data mining methods and utilities. Yogish and Raju [16] have presented an approach for clustering web users by using ART1 neural network based clustering algorithm. The performance of this method is compared with K-means and SOM clustering algorithms. Silva et al [20] presented a current survey on data stream clustering. Ichikawa and Morishita [31] have introduced a new and simple method with heuristic feature that reduces the computational time. Pattabiraman et al. [34] used three different clustering methods to cluster forum threads and discussed on the improvement of accuracy.

Parimala and Palanisamy [35] have introduced a new term "MFCC" i.e. Multitype Feature Co-selection for Clustering. To perform clustering of web documents, it exploits different type of feature classes. The authors also removed some challenges of search engine. Wang et al [37] has introduced AFS global K-means algorithm. In this method the distance based on AFS topology neighborhood is employed to determine initial cluster center. Execution time of K-means clustering algorithm has been reduced by Lee and Lin [57]. Sarma et al. [59] has proposed a prototype based hybrid approach to accelerate the traditional K-means clustering. Lam et al. [74] has proposed a PSO based K-means clustering for gene expression to enhance cluster matching. Sharma and Fotedar [44] have made a review on different data mining techniques used for software effort estimation.

Krey et al. [53] have presented order constrained solution in K-means as a more stable method for clustering of sound features. Huwang and Su [19] have improved the traditional K-means algorithm by making analysis on the statistical data. Xue and Liu [24] proposed a new approach to solve problems of clustering, which combines membrane computing with K-means algorithm.

IV. LIMITATIONS

K-means clustering has some of the limitations which need to get overcome. Several people got multiple limitations while working on their research with K-means algorithm. Some of the common limitations are discussed below.

Outliers

It has been observed by several researchers that, when the data contains outliers there will be a variation in the result that means no stable result from different executions on the same data. Outliers are such objects they present in dataset but do not result in the clusters formed. Outliers can also increase the sum of squared error within clusters. Hence it is very important to remove outliers from the dataset. Outliers can be removed by applying preprocessing techniques on original dataset [71].

Number of clusters

Determining the number of clusters in advance is always been a challenging task for K-means clustering approach. It is beneficial to determine the correct number of clusters in the beginning. It has been observed that sometimes the number of clusters are assigned according to the number of classes present in the dataset. Still it is an issue that on what basis the number of clusters should be assigned [49].

Empty clusters

If no points are allocated to a cluster during the assignment step, then the empty clusters occurs. It was an earlier problem with the traditional K-means clustering algorithm [71].

Non globular shapes and sizes

With the K-means clustering algorithm if the clusters are of different size, different densities and non globular shapes, then the results are not optimal. There is always an issue with the convex shapes of clusters formed [77].

V. APPLICATIONS

There are diverse applications of clustering techniques in the fields of finance, health care, telecommunication, scientific, World Wide Web, etc. Some of the applications are discussed below.

Clustering Algorithm in Identifying Cancerous Data

Clustering algorithm can be used in identifying the cancerous data record within a dataset. Different people tried on this application by assigning labels to known samples of datasets as cancerous and non-cancerous. Then randomly the data samples are mixed together and different clustering algorithms were applied. The result of clustering has been analyzed to know the correctly clustered samples. Accuracy of clustering is calculated easily as the labels of samples were known initially [41][72].

Clustering Algorithm in Search Engines

Clustering algorithm plays an important role in the functioning of search engines. Hence it will act as a backbone to search engines. Search engines try to group similar kind of objects into one cluster and dissimilar objects into other. The performance of the search engines depend on the working of the clustering techniques. The chances of getting the required results on the front page are more if the clustering technique is

better [64].

Clustering Algorithm in Academics

Students' academic progress monitoring has been a vital issue for academic society of higher learning. With clustering technique this issue can be managed easily. Based on the scores obtained by the students they are grouped into different clusters, where each cluster shows the different level of performance. By calculating the number of students' in each cluster we can determine the average performance of a class all together. [50].

Clustering Algorithm in Wireless Sensor Network based Application

Clustering Algorithm can be used efficiently in Wireless Sensor Network's based application. It can be used in landmine detection. Clustering algorithm plays a role of finding the cluster heads which collects all the data in its respective cluster. [3][28]

VI. CONCLUSION

In this paper, we have made a survey on work carried out by different researchers using K-means clustering approach. We also discussed the evolution, limitations and applications of K-means clustering algorithm. It is observed that a lot of improvement has been made to the working of K-means algorithm in the past years. Maximum work carried out on the improvement of efficiency and accuracy of the clusters. This field is always open for improvements. Setting appropriate initial number of clusters is always a challenging task. At the end it is concluded that although there has been made plenty of work on K-means clustering approach, there is a scope for future enhancement.

VII. REFERENCES-

[1] A. Kumar, R. Sinha, V. Bhattacherjee, D. S. Verma, S. Singh, "modeling using K-means clustering algorithm", *IEEE 2012, 1st international conference on recent advances in information technology(RAIT).*

[2] A. K. Jain, prof. S. Maheshwari, "survey of recent clustering techniques in data mining", *international journal of computer science and management research*, vol 1 issue 1 Aug 2012.

[3] A. Saurabh, A. Naik, "Wireless sensor network based adaptive landmine detection algorithm, " *2011 3rd International Conference on Electronics Computer Technology (ICECT)*, vol.1, no., pp.220, 224, 8-10 April 2011.

[4] B. Biggio, I. Pillai, S. R. Bulò, D. Ariu, M. Pelillo, F. Roli, "is data clustering in adversial settings secure?", *Proceedings of the 2013 ACM workshop on Artificial intelligence and security*, Pages 87-98.

[5] B. K. Mishra, N. R. Nayak, A. Rath, S. Swain, "far efficient K-means clustering algorithm", *Proceedings of the 2012 ACM's International Conference on Advances in Computing, Communications and Informatics,*

Pages 106-110.

[6] B. K Tripathy, A. Ghosh, G.K. Panda, "Kernel based K-means clustering using rough set, " *2012 International Conference on Computer Communication and Informatics (ICCCI)*, vol., no., pp.1, 5, 10-12 Jan. 2012.

[7] D. Jimenez, E. Ferretti, V. Vidal, P. Rosso, and C. F. Enguix, "The influence of semantics in IR using LSI and K-means clustering techniques", *ACM Proceedings of the 1st international symposium on Information and communication technologies*, Pages 279-284.

[8] D. Napoleon, P.G. Lakshmi, "An efficient K-Means clustering algorithm for reducing time complexity using uniform distribution data points, " *Trendz in Information Sciences & Computing (TISC), IEEE 2010*, vol., no., pp.42, 45, 17-19 Dec. 2010.

[9] D. Sisodia, L. Singh, S. Sisodia, K. Saxena, "Clustering Techniques: A Brief Survey of Different Clustering Algorithms", *International Journal of Latest Trends in Engineering and Technology (IJLTET)*, Vol. 1 Issue 3 September 2012.

[10] D. S. Modha, W. S. Spangler, "*Feature Weighting in k-Means Clustering*", *ACM Journal of Machine Learning*, Volume 52 Issue 3, September 2003, Pages 217 – 237.

[11] D. T. Pham, S. S. Dimov, and C. D. Nguyen, "Selection of K in K-means clustering", *Proc. IMechE Vol. 219 Part C: J. Mechanical Engineering Science, IMechE 2005.*

[12] E. A. Khadem, E. F. Nezhad, M. Sharifi, "Data Mining: Methods & Utilities", *Researcher*2013; 5(12):47-59. (ISSN: 1553-9865).

[13] E. Casper, C. C. Hung, E. Jung, and M. Yang, "**A quantum-modeled K-means clustering algorithm for multi-band image segmentation**", *Proceedings of the 2012 ACM Research in Applied Computation*, Pages 158-163.

[14] E. Kijsipongse, S. U-ruekolan, "Dynamic load balancing on GPU clusters for large-scale K-Means clustering, " *2012 IEEE International Joint Conference on Computer Science and Software Engineering (JCSSE)*, vol., no., pp.346, 350, May 30 2012-June 1 2012.

[15] G. D. Fatta, F. Blasa, S. Cafiero, G. Fortino, "Epidemic K-Means Clustering, " *2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW)*, vol., no., pp.151, 158, 11-11 Dec. 2011.

[16] H.K. Yogish, G.T. Raju, "Clustering of Preprocessed Web Usage Data Using ART1 Neural Network and Comparative Analysis of ART1, K-Means and SOM Clustering Techniques, " *2013 5th IEEE International Conference on Computational Intelligence and Communication Networks (CICN)*, vol., no., pp.322, 326, 27-29 Sept. 2013.

[17] H.T Dashti, T Simas, R.A Ribeiro, A Assadi, A Moitinho, "MK-means-Modified K-means clustering algorithm, " *The 2010 IEEE International Joint Conference on Neural Networks (IJCNN)*, vol., no., pp.1, 6, 18-23 July 2010.

[18] H. Xiong; J. Wu; J. Chen, "K-Means Clustering Versus Validation Measures: A Data-Distribution Perspective, " *IEEE Transactions on Systems, Man, and*

Cybernetics, Part B: Cybernetics, vol.39, no.2, pp.318, 331, April 2009.

[19] H. Xiuchang, SU Wei, "An Improved K-means Clustering Algorithm", *JOURNAL OF NETWORKS*, VOL. 9, NO. 1, JANUARY 2014.

[20] J. A. Silva, E. R. Faria, R. C. Barros, E. R. Hruschka, A. C. P. L. F. d. Carvalho, J. Gama, "Data stream clustering: A survey", *ACM Computing Surveys (CSUR)*, Volume 46 Issue 1, October 2013, Article No. 13.

[21] J. Feng; Z. Lu; P. Yang; X. Xu, "A K-means clustering algorithm based on the maximum triangle rule, " *2012 IEEE International Conference on Mechatronics and Automation (ICMA)*, vol., no., pp.1456, 1461, 5-8 Aug. 2012.

[22] J. Wang; J. Wang; Q. Ke; G. Zeng; S. Li, "Fast approximate k-means via cluster closures, " *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol., no., pp.3037, 3044, 16-21 June 2012.

[23] J. Wang; X. Su, "An improved K-Means clustering algorithm, " *2011 IEEE 3rd International Conference on Communication Software and Networks (ICCSN)*, vol., no., pp.44, 46, 27-29 May 2011.

[24] J. Xue, X. Liu, "A K-nearest Based Clustering Algorithm by P Systems with Active Membranes", *Journal of Software*, Vol 9, No 3 (2014), 716-725, Mar 2014.

[25] J. Zhu; H. Wang, "An improved K-means clustering algorithm, " *2010 The 2nd IEEE International Conference on Information Management and Engineering (ICIME)*, vol., no., pp.190, 192, 16-18 April 2010.

[26] J. MacQueen, "Some methods for classification and analysis of multivariate observations." *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1: Statistics, 281--297, University of California Press, Berkeley, Calif., 1967.

[27] K. Alsabti, S. Ranka, V. Singh, "An efficient k-means clustering algorithm", *1998 Proceedings of IPPS/SPDP Workshop on High Performance Data Mining.(date location).*

[28] K. Akkaya, F. Senel, B. McLaughlan, "Clustering of Wireless Sensor and Actor Networks based on Sensor Distribution and Inter-actor Connectivity" *ACM Journal of Parallel and Distributed Computing*, volume 69, Issue 6, June, 2009, Pages 573-587.

[29] K. Honda, R. Nonoguchi, A. Notsu, H. Ichihashi, "PCA-guided k-Means clustering with incomplete data, " *2011 IEEE International Conference on Fuzzy Systems (FUZZ)*, vol., no., pp.1710, 1714, 27-30 June 2011.

[30] K. Hornik, I. Feinerer, M. Kober, C. Buchta, "Spherical k-Means Clustering", *Journal of Statistical Software*, September 2012, Volume 50, Issue 10.

[31] K. Ichikawa, S. Morishita, "A simple but powerful heuristic method for accelerating k-means clustering of large-scale data in life science, " *2013 IEEE Transactions on Computational Biology and Bioinformatics*, vol.PP, no.99, pp.1, 1.

[32] K. Murugesan, J. Zhang, "Hybrid Bisect K-Means Clustering Algorithm, " *2011 International Conference on Business Computing and Global Informatization (BCGIN)*, vol., no., pp.216, 219, 29-31 July 2011.

[33] K. NIGAM, A. K. MCCALLUM, S. THRUN, T. MITCHELL, "Text Classification from Labeled and Unlabeled Documents using EM", *ACM journal of Machine Learning-Special issue on information retrieval* 1999.

[34] K. Pattabiraman, P. Sondhi, C. Zhai, "exploiting forum thread structures to improve clustering", *ACM Proceedings of the 2013 Conference on the Theory of Information Retrieval*, Pages 15.

[35] K. Parimala, Dr. V. Palanisamy, " Enhanced Performance of Search Engine with Multitype Feature Co-Selection of K-Means Clustering Algorithm ", *IJASCSE*, Vol 2, Issue 1, 2013.

[36] K. Wagsta, C. Cardie, S. Rogers, S. Schroedl, "Constrained K-means Clustering with Background Knowledge", *ACM Proceedings of the Eighteenth International Conference on Machine Learning*, 2001, p. 577-584.

[37] L. Wang, X. Liu, Y. Mu, "The Global k-Means Clustering Analysis Based on Multi-Granulations Nearness Neighborhood", *Mathematics in Computer Science*, March 2013, Volume 7, Issue 1, pp 113-124.

[38] L. Xinwu, "Research on text clustering algorithm based on improved K-means, " *2010 International Conference on Computer Design and Applications (ICCDA)*, vol.4, no., pp.V4-573, V4-576, 25-27 June 2010.

[39] M. Bakhshi, M.R.F. Derakhshi, *E. Zafarani*, "Review and Comparison between Clustering Algorithms with Duplicate Entities Detection Purpose", *International Journal of Computer Science Emerging Technology*, Vol-3, No 3, June, 2012.

[40] M. Estlick, M. Leeser, J. Theiler, J.J. Szymanski, "algorithmic transformations in the implementation of K-means clustering on reconfigurable hardware", *Proceedings of the 2001 ACM/SIGDA ninth international symposium on Field programmable gate arrays*, Pages 103-110.

[41] M. Girolami, C. He, "Probability density estimation from optimally condensed data samples, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.25, no.10, pp.1253, 1264, Oct. 2003.

[42] M.M.T. Chiang, B. Mirkin, "Intelligent Choice of the Number of Clusters in *K*-Means Clustering: An Experimental Study with Different Cluster Spreads", *Journal of Classification*, March 2010, Volume 27, Issue 1, pp 3-40.

[43] M. Sathya, J. Jayanthi, N. Basker, "Link based K-Means clustering algorithm for information retrieval, " *2011 IEEE International Conference on Recent Trends in Information Technology (ICRTIT)*, vol., no., pp.1111, 1115, 3-5 June 2011.

[44] M. Sharma, N. Fotedar, "Software Effort Estimation with Data Mining Techniques-A Review", *International journal of engineering sciences and research technology*, 3(3): March, 2014, ISSN: 2277-9655.

[45] M.S. Mahmud, M.M. Rahman, M.N. Akhtar, "Improvement of K-means clustering algorithm with better initial centroids based on weighted average, " *2012 7th IEEE International Conference on Electrical & Computer Engineering (ICECE)*, vol., no., pp.647, 650, 20-22 Dec. 2012

[46] M. Vijayalakshmi, M.R. Devi, " A Survey of Different Issue of Different clustering Algorithms Used in Large Data sets", *International Journal of*

Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 3, March 2012 ISSN: 2277 128X.

[47] N. Cravino, J. Devezas, Á. Figueira, "Using the overlapping community structure of a network of tags to improve text clustering", *Proceedings of the 23rd ACM conference on Hypertext and social media*, Pages 239-244.

[48] N. T. Son; D. T. Anh, "Two Different Methods for Initialization the I-k-Means Clustering of Time Series Data", *2011 Third International Conference on Knowledge and Systems Engineering (KSE)*, vol., no., pp.3, 10, 14-17 Oct. 2011.

[49] O. A. Abbas, "comparisons between data clustering algorithms", *the international Arab journal of information technology*, vol. 5, no. 3, July 2008.

[50] O. J Oyelade, O. O Oladipupo, I. C Obagbuwa, "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", *(IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, _o. 1, 2010.*

[51] O. Shamir, N. Tishby, "Stability and model selection in *k*-means clustering", *Machine Learning*, September 2010, Volume 80, Issue 2-3, pp 213-243.

[52] R. Vij, S. Kumar, "Improved k-means clustering algorithm for two dimensional data", *ACM Proceedings of the Second International Conference on Computational Science, Engineering and Information Technology*, Pages 665-670.

[53] S. Krey, U. Ligges, F. Leisch, "Music and timbre segmentation by recursive constrained *K*-means clustering", *Computational Statistics*, February 2014, Volume 29, Issue 1-2, pp 37-50.

[54] S. Na; L. Xumin; G. Yong, "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm, " *2010 Third IEEE International Symposium on Intelligent Information Technology and Security Informatics (IITSI)*, vol., no., pp.63, 67, 2-4 April 2010.

[55] S. Ren; A. Fan, "K-means clustering algorithm based on coefficient of variation, " *2011 4th International Congress on Image and Signal Processing (CISP)*, vol.4, no., pp.2076, 2079, 15-17 Oct. 2011.

[56] S.S. Lee, J.C. Lin, "An accelerated *K*-means clustering algorithm using selection and erasure rules", *Journal of Zhejiang University SCIENCE C*, October 2012, Volume 13, Issue 10, pp 761-768.

[57] S.S. Lee, J.C. Lin, "Fast K-means clustering using deletion by center displacement and norms product (CDNP)", *Journal of Pattern Recognition and Image Analysis*, Volume 23 Issue 2, April 2013, Pages 199-206.

[58] T.A Runkler, "Partially supervised k-harmonic means clustering, " *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, vol., no., pp.96, 103, 11-15 April 2011.

[59] T. H. Sarma, P. Viswanath, B. E. Reddy, "A hybrid approach to speed-up the *k*-means clustering method", *International Journal of Machine Learning and Cybernetics*, April 2013, Volume 4, Issue 2, pp 107-117.

[60] T.H Sarma, P. Viswanath, B.E. Reddy, "A fast approximate kernel k-means clustering method for large data sets, " *Recent Advances in Intelligent*

Computational Systems (RAICS), 2011 IEEE, vol., no., pp.545, 550, 22-24 Sept. 2011.

[61] T. Jinlan, Z. Lin, Z. Suqin, L. Lu, "Improvement and Parallelism of *k*-Means Clustering Algorithm", *Tsinghua Science & Technology*, Volume 10, Issue 3, June 2005, Pages 277–281.

[62] T. Kanungo, D. M Mount, N.S. Netanyahu, C.D. Piatko, R. Silverman, A. Y. Wu, "An efficient k-means clustering algorithm: analysis and implementation, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.24, no.7, pp.881, 892, Jul 2002.

[63] T. Li; Y. Chen; X. Mu; M. Yang, "An improved fuzzy k-means clustering with k-center initialization, " *2010 Third IEEE International Workshop on Advanced Computational Intelligence (IWACI)*, vol., no., pp.157, 161, 25-27 Aug. 2010.

[64] T. Liu, C. Rosenberg, H.A. Rowley, "Clustering Billions of Images with Large Scale Nearest Neighbor Search, " *Applications of Computer Vision, 2007. IEEE Workshop on WACV '07*, vol., no., pp.28, 28, Feb. 2007.

[65] V.V. Vu, N. Labroche, B.B. Meunier, "Active Learning for Semi-Supervised K-Means Clustering, " *2010 22nd IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, vol.1, no., pp.12, 15, 27-29 Oct. 2010.

[66] W. Min; Y. Siqing, "Improved K-means clustering based on genetic algorithm, " *2010 IEEE International Conference on Computer Application and System Modeling (ICCASM)*, vol.6, no., pp.V6-636, V6-639, 22-24 Oct. 2010.

[67] W. Yintong; L. Wanlong; G. Rujia, "An improved k-means clustering algorithm, " *World Automation Congress (WAC), 2012*, vol., no., pp.1, 3, 24-28 June 2012.

[68] X. Li; J. Yang; Q. Wang; J. Fan; P. Liu, "Research and Application of Improved K-Means Algorithm Based on Fuzzy Feature Selection, " *FSKD '08. Fifth International Conference on Fuzzy Systems and Knowledge Discovery*. vol.1, no., pp.401, 405, 18-20 Oct. 2008.

[69] X. Wang, "A fast exact k-nearest neighbors algorithm for high dimensional search using k-means clustering and triangle inequality", *The 2011 International Joint Conference on Neural Networks (IJCNN),*, vol., no., pp.1293, 1299, July 31 2011-Aug. 5 2011.

[70] X. Wu, C. Yao, "Application of improved K-means clustering algorithm in transit data collection, " *2010 3rd International Conference on Biomedical Engineering and Informatics (BMEI)*, vol.7, no., pp.3028, 3030, 16-18 Oct. 2010.

[71] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.H. Zhou, M. Steinbach, D.J. Hand, D. Steinberg, "Top 10 algorithms in data mining", *Knowledge and Information Systems*, January 2008, Volume 14, Issue 1, pp 1-37.

[72] X. Y. Wang, J. M. Garibaldi. "A comparison of fuzzy and non-fuzzy clustering techniques in cancer diagnosis", Proceedings *of the 2nd International Conference on Computational Intelligence in Medicine and*

Healthcare, 2005.

[73] Y. Cheng, Y. Xie, K. Zhang, A. Agrawal, A. Choudhary, "CluChunk: clustering large scale user-generated content incorporating chunklet information", *ACM Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems,* Programming Models and Applications, Pages 12-19.

[74] Y.K. Lam, P. W. M. Tsang, C.S. Leung, "PSO-based K-Means clustering with enhanced cluster matching for gene expression data", *Neural Computing and Applications*, June 2013, Volume 22, Issue 7-8, pp 1349-1355.

[75] Y.M. Cheung, "k∗-Means: A new generalized *k*-means clustering algorithm", *Pattern Recognition Letters, Volume 24, Issue 15, November 2003, Pages 2883-2893.*

[76] Y. Sun; G. Liu; K. Xu, "A k-Means-Based Projected Clustering Algorithm, " *2010 Third International Joint Conference on Computational Science and Optimization (CSO)*, vol.1, no., pp.466, 470, 28-31 May 2010.

[77] Y. S. Patil, M.B. Vaidya, " A Technical Survey on cluster analysis in data mining", *International Journal of Emerging Technology and Advanced Engineering*, ISSN 2250-2459, Volume 2, Issue 9, September 2012.

[78] Y. Zhang, H. Shi, D. Zhang, "K-means clustering based on self-adaptive weight, " *2012 2nd International Conference on Computer Science and Network Technology (ICCSNT)*, vol., no., pp.1540, 1544, 29-31 Dec. 2012.

[79] Y. Zhu, J. Yu, C. Jia, "Initializing K-means Clustering Using Affinity Propagation, " *Ninth International Conference on Hybrid Intelligent Systems, 2009. HIS '09.* vol.1, no., pp.338, 343, 12-14 Aug. 2009.