International Journal of Engineering Research and Advanced Technology (IJERAT)

DOI: 10.31695/IJERAT.2020.3646

E-ISSN : 2454-6135

Volume.6, Issue 9
September -2020

Subject Review: Text Clustering Algorithms

Zuhair Hussein Ali[1], Amal Abbas Kadhim[2] and Azal Minshed Abid[3]

Department of Computer Science

[1-3]Mustansiriyah University, College of Education

Iraq

ABSTRACT

Clustering algorithms are taking attention in recent times, according to a huge amount of datasets and the growth of parallelized computing architectures. The goal of clustering algorithms is to divided the dataset into clusters, such that objects within the same cluster are similar to each other and differ from objects of other clusters. Clustering algorithms play an important role in information retrieval, indexing and text summarization. In this paper a brief overview of several clustering algorithms is discussed

Key Words:*Clustering, Hierarchical, Density based, Partition based,Grid based*.

1. INTRODUCTION

The motivation behind clustering algorithms is to make sense of and extract value from huge amount of structured and unstructured data. In case of working with the tremendous volumes of unstructured information, it just attempt to segment the data into some sort of logical grouping before trying to investigate it. Clustering can be defined as a gathering of comparative objects into a set known as clusters. Objects in a single cluster are probably going to be diverse when contrasted with objects gathered under another cluster[1]. Clusters can also be defined as sets of data points that share similar attributes, and clustering algorithms are the methods that group these data points into different clusters based on their similarities. Clustering is one of the principle undertakings in exploratory data mining and is likewise a strategy utilized in statistical data analysis[2]. High dimensionality is the major problem in document clustering, this problem can be solved using efficient algorithms for documents clustering, While clustering isn't one explicit algorithm, it is a general assignment that can be comprehended by many algorithms. Some of the well known clustering techniques that are utilized incorporate hierarchical, partitioning, density-based and model-based. Clustering can also be defined as groups of data points that share similar characteristics, and clustering algorithms are methods that group these data points into different groups according to their similarities. Text document of relevant topics can be grouped together using clustering algorithms. Text clustering used efficiently in text summarization and information retrieval [3].

In this paper different algorithm of documents clustering are presents that help researchers choose which algorithm is best for document clustering dependent on the prerequisites.

2. CLUSTERING ALGORITHMS

Figure 1 presents the main algorithms for big data clustering, that consists of hierarchical algorithms, partition based algorithms, density based algorithms and grid based algorithms.

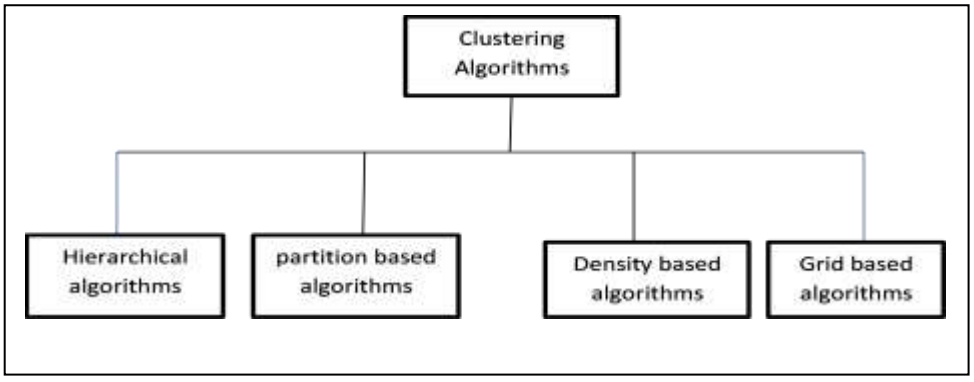

Figure 1: Clustering Algorithm.

2.1 Hierarchical Clustering Algorithms

Hierarchical clustering methods are of various idea from the sequential algorithms, hierarchical algorithms produce a hierarchy of clusters instead single cluster. Hierarchical clustering is one of the most widely used methods of text clustering. This is due to its high efficiency, as it relies on generating a hierarchical nested category, in other words, any change of hierarchical category hierarchical, corresponding object also will be changed. The result of one tree for hierarchical grouping can produce multiple nodes. These nodes can in turn produce multiple nodes, and so on[4].

Hierarchical clustering contains two different strategies. The first strategy based on starting from the top to the bottom of the tree called top-bottom, while the second strategy starts from the bottom to the top based on merging two similar clusters, This process repeated until the top node is reached[5].

Top-bottom clustering strategy based on dividing the document's text into a number of disjoint points, so that each similar sub points are gathered in one the cluster. Similarity is implemented as follows: choose an appropriate centroid point, compute the distance between every point and the centroid point, select the points that are closer to the centroid to be in the same cluster. While bottom-up strategy based on starting from a single object that merged with another similar object to create a new cluster. This process repeated for new generated clusters and stopped when there is no more cluster can be generated[6]. BIRCH and CURE are some of well-known algorithms for this clustering technique.

2.1.1 BIRCH Clustering Algorithm

BIRCH Acronym for *Balanced Iterative Reducing and Clustering Using Hierarchies.* Is an unsupervised clustering algorithm. The BIRCH algorithm is more appropriate for the situation where the quantity of information is huge and the number of categories K is moderately enormous. The BIRCH algorithm is very fast because it scans the data set only once for the purpose of clustering. It uses a tree structure to produce a cluster. It is commonly named the Clustering Feature Tree (CF Tree). Each node of this tree consists of various Clustering features. Three important parameters are required for the BIRCH algorithm: the threshold, the branching factor, and the number of clusters [7].

2.1.2 CURE Clustering Algorithm

CURE stands for Clustering Using Representatives is an efficient algorithm for huge data clustering. Rather than utilizing one point centroid, as in a large portion of data mining clustering algorithms. CURE utilizes a lot of all around representative points, for efficiently managing the clusters and removing the outliers and also for organizing the spherical and non-spherical clusters. It is helpful for finding gatherings and distinguishing interesting distributions in the underlying data. The CURE algorithm is mainly based on random sampling and partitioning. Random samples are taken from the data set and partition each partition, then included in the sub cluster, that in turn cluster in the second pass this process is repeated for all data points in the data set[8].

2.2 Partition Based Clustering

The basic idea of these algorithms is based on predetermining a number of clustering. Given a data set of n objects, Each object must have its place to exactly one cluster and every cluster contains at least one object. The main idea behind a partition algorithm based on spread the data set objects into partition, where each partition represent a cluster. The weakness of such algorithms is that the point may be further from the center similar to the distance of the point from another center, so it may happen that the point belongs to the incorrect cluster. K-means algorithm and Fuzzy c-means algorithm are some of well-known algorithms for this clustering technique [9].

2.2.1 K-means Clustering Algorithm

Introduced in 1967 by James Macqueen. The basic idea based on considering the centroid as a cluster. The mean of points within the same cluster used to compute the centroid. The distance between the point and the centroid used as an objective function to determine the affiliation of point to the cluster. The algorithm consists of many steps. The first step involves the selection of K clusters from the data set as the initial centers, While the second step involves assigning each point in the data set to the one cluster based on computing the distance between the point and the cluster center the point assign to the nearest center. The final step includes recomputing of the mean for each cluster. This process repeated until no change occurred. The K-mean algorithm very efficient for large data sets, faster and provide tighter clustering than hierarchical algorithm. But it suffers from the difficulty to predict the k clusters and does not work well with global clusters[10].

DOI : 10.31695/IJERAT.2020.3646

2.2.2 Fuzzy C-means Clustering Algorithm

In the partition clustering algorithm the object belongs to only one cluster and There is no possibility of it being present in another cluster. Fuzzy c-means overcoming this limitation by the probability of assigning the object to more than one cluster by using a membership function. Fuzzy clustering used widely in pattern recognition and marketing.

Fuzzy c-means is an extension of k-means. As mentioned previously, each object assigned to one cluster in the k-means algorithm, while in the fuzzy c-means each object in the dataset assigned to more than one cluster. Each object has a degree of membership of belonging to each cluster. Fuzzy c-means useful for huge data and always converges occurred, while it suffers from long computational time, sensitive to initial values and sensitive to noise [11].

2.3 Density Based Clustering Algorithm

Density based algorithm proposes that, connected objects with the same density must belong to the same category. Density based algorithm start from initial object and expand in any direction based on the same density. Therefore, can determine the arbitrary shape, also the algorithm work well with the existing of the noise and can efficiently detect outliers objects. DBSCAN algorithm and OPTICS algorithm are some of well-known algorithms for this clustering technique [12].

2.3.1 DBSCAN Clustering Algorithm

The main idea of DBSCAN algorithm based on calculating the number of points in a fixed- radius neighborhood, then determined that these points are connected if they exist in one another's neighborhood. So there are two parameters in this algorithm Eps-neighborhood and minimum points. The Eps-neighborhood determine the distance to decide whether the point belongs to same cluster or not, while the minimum points determined the minimum number of points in Eps-neighborhood. These two parameters specified by the users. Also, there are two important points must be taken in the consideration: core point and border points. The core point is the point that has a larger value than other points that exist within the Eps-neighborhood. Border point is the point that has less value than a precise number of points that is Minimum points. DBSCAN algorithm useful to detect arbitrary shaped clusters also can handle the noise and outlier efficiently, But Cannot perform well with large differences in densities and Not suitable when various densities involve [13].

2.3.2 OPTICS Clustering Algorithm

OPTICS stand for **Ordering points to identify the clustering structure**, used to find density cluster in spatial data. OPTICS algorithm similar to DBSCAN In terms of the existence of Eps-neighborhood, minimum points, core point and border point, The only difference is that it does not assign cluster memberships but stores the order in which the points are processed. This differences allows to overcome the weakness of the DBSCAN algorithm for detecting the varying density in the dataset. The main idea of the OPTICS algorithm based on ordering the dataset point such that similar points come to be neighbors in the ordering. Moreover, the density stored at every point as a distance that represents the acceptability of two points to be within the same clustering [14].

2.4 Grid Based Clustering Algorithm

The grid base algorithm is popular for handling large dataset. The main idea of the grid approach based on the value space surrounds the data point and not on its values. The basic steps of the algorithm based on partitioning data points into a number of grids, then calculating the cluster center using mean of data points, and finally investigate the neighbor cells. The main advantage of grid algorithm is very fast as compared with others clustering algorithm especially for large datasets. CLUQUE algorithm and STING algorithm are some of well-known algorithms for this clustering technique [15].

2.4.1 CLUQUE Clustering Algorithm

CLUQUE algorithm based on separating the vertices of a graph into different clusters, where each a clique in the graph represent a cluster. The basic procedure of the CLUQUE algorithm based on partition the data set points on cells and compute the number of points in each cell. A priori principle used to identify subspaces that contain clusters, and finally these subspaces used to detect a cluster. The main advantages of The CLUQUE are working efficiently with large dataset point and not sensitive to the order of dataset points, while the weakness of the algorithm the quality of the result based on the number of partitions used [16].

2.4.2 STING Clustering Algorithm

STING stands for STatistical Information Grid. The main idea of STING algorithm based on dividing the dataset into a hierarchy structure. At first starting from the root then the high level cell is partitioned into several low level cells. The statistical information stored in each cell. The statistical information include mean, minimum and maximum values. The STING algorithm is useful when there is large dataset [17].

DOI : 10.31695/IJERAT.2020.3646

3. CONCLUSIONS

This paper has presented a survey of clustering algorithms. These algorithms include hierarchal algorithms, partitioning algorithms, density based algorithms and grid based algorithm. The survey showed that hierarchal and partition algorithms are easy to implement and give accurate clustering when objects of the dataset are not similar in the density and be adjacent in the place, whereas density based and grid based algorithms are faster and perform well when there is different in the object density.

ACKNOWLEDGMENT

The authors would like to thank Mustansiriyah University (www.uomustansiriyah.edu.iq) Baghdad – Iraq for its support.

REFERENCES

1. Boris,L., AnaKosa,r., Bersant,D., Dženan,S., Peter,R. & Axel,K. (2018). Variations on the Clustering Algorithm BIRCH,vol 11 pp.44-53
2. Yadav , C., Wang , S. & Kumar, M20(13). Algorithms and approaches to handle large data sets - A survey. International Journal of Computer Science and Network. , vol 2,no.3,pp.1–5.

3 Aggarwal, C. & Zhai, C.(2012) A survey of text clustering algorithms Mining Text Data. New York, NY, USA. Springer-Verlag.p. 77–128.

4. Marjan,K.,uchaki, R., Zahra, A., Nasibeh, E, (2012), A survey of hierarchical clustering algorithms, The Journal of Mathematics and Computer Science,Vol. 5,No.3, pp.229- 240.
5. Chris, d.,& Xiaofeng. He. (2002), Cluster Merging And Splitting In Hierarchical Clustering Algorithms.
6. Contreras, P., & Murtagh, F.(2010). Fast hierarchical clustering from the Baire distance. In Classification as a Tool for Research, eds. H. Hocarek-Junge and C. Weihs, Springer, Berlin, pp.235–243.
7. Park H, Park J & Kwon ,Y.(2015). Topic clustering from selected area papers. Indian Journal of Science and Technology Oct;Vol. 8,No.26,PP:1–7.
8. Majumdar J., Udandakar S., Mamatha Bai B.G. (2019) Implementation of Cure Clustering Algorithm for Video Summarization and Healthcare Applications in Big Data. In: Shetty N., Patnaik L., Nagaraj H., Hamsavath P., Nalini N. (eds) Emerging Research in Computing, Information, Communication and Applications. Advances in Intelligent Systems and Computing, vol 906. Springer, Singapore. https://doi.org/10.1007/978-981-13-6001-5_46.

9. Liuzzi, G., Lucidi, S. & Piccialli, V. (2010). A partition-based global optimization algorithm. J. Global Optimization. Vol.48. PP.113-128. 10.1007/s10898-009-9515-y.
10. BOTTOU, L. & BENGIO, Y. 1995. Convergence properties of the K- means algorithms.Advances in Neural Information Processing Systems Vol. 7, PP. 585-592.
11. Said, E., Rowayda ,A., Mohamed ,A. (2015). An efficient brain mass detection with adaptive clustered based fuzzy C-mean and thresholding. 2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA):PP. 429–433.
12. Fahad ,A., Alshatri, N., Tari, Z., Alamri, A.(2014). A survey of clustering algorithms for Big Data: Taxonomy and empirical analysis. IEEE Transactions on Emerging Topics in Computing. Vol. 2,No. 3,PP.:267–79.
13. Campello, R., Moulavi, D.,Zimek, A.& Sander, J. (2013). A framework for semi-supervised and unsupervised optimal extraction of clusters from hierarchies. Data Mining and Knowledge Discovery.Vol. 27,No. 3,PP, 344. doi:10.1007/s10618-013-0311-4.
14. Achtert, E., Böhm, C.; Kriegel, H. P., Kröger, P. ,Müller-Gorman, I.& Zimek, A. (2006). Finding Hierarchies of Subspace Clusters. LNCS: Knowledge Discovery in Databases: PKDD 2006. Lecture Notes in Computer Science. pp. 446–453.
15. MR, ILANGO & MOHAN, Dr. (2010). A Survey of Grid Based Clustering Algorithms. International Journal of Engineering Science and Technology.

DOI : 10.31695/IJERAT.2020.3646

16. Park ,H., Park .J., Kwon YB.(2015). Topic clustering from selected area papers. Indian Journal of Science and Technology. Vol. 8,No.26, PP.1–7.
17. Hans,P., & Arthur Z.,(2010) ,Subspace clustering, Ensemble clustering, Alternative clustering, Multiview clustering: What can we learn from each other, In Proc. 1st Int'l workshop on discovering, summarizing and using multiple clusterings.

DOI : 10.31695/IJERAT.2020.3646