Review of Clustering Techniques for Finding the Similarity in Articles

Usha Rani
M.Tech Scholar,
Department of
Computer Science & engineering,
Ajay Kumar Garg
Engineering College,
Ghaziabad, India

Shashank Sahu
Associate Professor,
Department of Computer
Science & Engineering
Ajay Kumar Garg
Engineering College,
Ghaziabad, India

ABSTRACT
Clustering is an important technique in data mining. It is a technique in which grouping of item taken place into the clusters in such a way that items of same cluster have more similarity than the items into another cluster, but is very dissimilar to the item in other clusters. The aim of document clustering is to make a set of clusters of given documents in such a way that document of each cluster have more similarity than the documents of other clusters. This paper reviews various techniques of clustering which can be divided mainly into two groups that are hierarchical and partitional clustering.

Keywords
Clustering, Hierarchical clustering, Partitional clustering.

1. INTRODUCTION
Clustering is a method that arranges the large number of unordered text documents into a small number of valid clusters. Clustering can be done by two ways: Partitioning method and Hierarchical method. Hierarchical clustering group the items into a tree of cluster, it is also known as hierarchical cluster analysis. Hierarchical clustering method is classified further under two categories: Divisive hierarchical clustering and agglomerative hierarchical clustering. This depends on how the hierarchical decomposition takes place [1]. Agglomerative approach is also called bottom up approach. It start with each object as a separate cluster in starting then merging the pairs of cluster as proceed up into the hierarchy. The process continues until all the clusters are merged into one cluster or a termination condition meet. Divisive approach is also known as top down approach. It works as agglomerative approach but in opposite direction. It starts by taking all the objects as one single cluster, at each next step it divides or break the clusters into smaller clusters until a termination condition reached or each object fall into separate cluster

[2].Advantage of hierarchical clustering**:**
Applicable to any attribute type.

- It is easy to handle any type of similarity.
- No need to provide number of cluster in advance.
- Embedded flexibility in respect of level of granularity.
- More suitable for problems having point linkage.

Disadvantage of hierarchical clustering**:**

- Uncertainty regarding termination condition.
- Once a merging or splitting is taken then this is not reversible or cannot be undone.
- Degradation in high dimensional spaces.
- High in time complexity
- Very expensive for high dimensional and huge dataset [3].

2. HIERARCHICAL CLUSTERING
It is a way of cluster analysis which builds a hierarchy of clusters. The quality of hierarchical clustering suffers from its inability to perform reverse action, once a merge or split decision has been taken. One possible way for improving hierarchical clustering quality is to combine it with other clustering techniques these are called improved hierarchical clustering algorithms.

2.1 Birch
Its mean balanced iterative reducing and clustering using hierarchies. It is designed for performing clustering on large amount of numeric data. It performs hierarchical clustering at initial stage and then later stage iterative partitioning. It solves two problems in agglomerative clustering method weakness to undo what was done in earlier step and scalability problem. This methods are introduced in birch (1) clustering feature(CF) and (2) clustering feature tree(CF tree).These method are used for analysis of cluster representations. These help in big database to gain scalability and good speed. It is multiphase clustering technique. First phase is for overlook which is used as good clustering then at next phase quality of clustering is improved [4].

2.2 Chemeleon
It is an agglomerative hierarchical clustering algorithm and it is based on the concept of nearest-neighbor graph, an edge is removed if both vertices don't fall in the closest points related to each other. It overcomes the problem of agglomerative hierarchical clustering algorithms that these methods do not use of information about the nature of individual clusters being used [5]. It uses the dynamic modeling and measure the similarity of clusters based on this modeling. It is based on two phases:

- A graph partitioning algorithm is used to partition the large cluster into smaller sub clusters.
- Repeatedly merging of sub clusters which are coming from the previous step to obtain a genuine cluster.

It can make clusters of distinct shapes and sizes.

3. PARTITIONAL CLUSTERING

In partitional clustering the objects are divided into k partition, where each partition represents a cluster. Each partition contains at least one object and every object belongs to only one cluster. Each partition is formed based on some objective functions such as minimizing the square error [6]. Advantage of partitional clustering:

- Best for datasets having compact spherical clusters that are well-separated.
- It permit object to exit from cluster to another cluster to improve criterion.
- More scalable, reliable and simple approach.

Disadvantage of partitional clustering:

- In high dimensional space effectiveness degradation takes place.
- Very sensitivity to initialization phase.
- Lack to deal with non-convex clusters of different size and density.
- User has to specify number of cluster in advance.

Cluster are described always with a central vector .It is not necessary that it may be a member of data set. The algorithm always assigns the item to the nearest center. The centric based algorithm is K-means, CLARANS, K-medoids etc.

3.1 K-means

The term "*k*-means" was first used in 1967 by James MacQueen It is a smooth unsupervised learning algorithm. The process is very simple which classify a given data set into a number of clusters where the number of cluster is given in advance. First of all k centroids are defined, one for each cluster. These centroids must be placed in a smart way because different location shows different results [11]. The preferred choice is to place them as far as possible from each other. At next step each point in the data set is taken and associates it with nearest centroid. When there is no point left into data set then this step is completed. Again re-calculate the new k centroids and a new binding is done between the new centroid and data set a loop is formed because of this k centroid changes their location until no more changes are done .This aims so that objective function should be minimized or squared error function be minimized. The objective function is given below

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and cluster centre c_j, is an indicator of the distance of the given n data points from their respective cluster centres.

3.2 K-medoids

This method is partitional (dividing the dataset into groups) and first proposed in 1987.It is related to k-means but it uses medoids rather than mean/centroid .Medoid is a member of data set whose average dissimilarity to all the other members of the data set is minimal. It is the most central item of the data set. It begins by randomly selecting k items from the data set as initial medoids to represent the k clusters. All the remaining items are put into cluster whose medoids is closer to them. At next step a new medoids is determined which will represent the cluster in a better way after that again all the remaining items are put into cluster whose medoids is closer to them. In each iteration medoids change their location. All the step continues until no medoids changes its position. At the end all the data points from the data set are put into cluster based on nearest medoids. Number of cluster and data set is given as input and k clusters that minimize the sum of the dissimilarities of all the objects to their nearest medoids are the output.

$Z = \sum k\ i\text{-}1 \sum |x\text{-}mi|$

Z: Sum of absolute error for all data points in the data set

x: the data point in the data set

mi: is the medoid of cluster Ci

3.3 Clarans

Its mean clustering large applications based upon randomized search. It selects randomly k items from the data set which are used as current medoids. Next select an item y which is not current medoids and one medoids x, then find out whether replacing x with y improve absolute error, if yes then replacement take place otherwise not.

4. DENSITY BASED CLUSTERING

There is also third types of clustering techniques which is called density based clustering. In density based clustering, clusters are characterized as areas of higher density than the rest of the data set. Items in the scattered area that are required to separate clusters are considered noise and boundary points. It is in sensitive to the ordering of database and need two parameters. The quality of clustering depends on the distance measure which is used in the function. There is no need to tell the number of cluster in advance [7]. These methods are developed for making clusters of arbitrary shape, which may not necessarily of convex shape. These are unsuitable for high dimensional dataset [8]. The main density based clustering algorithms are DBSCAN and OPTICS.DBSCAN mean density based spatial clustering of applications with noise and OPTICS mean ordering points to identify the clustering structure. It is same as DBSCAN but it gives augmented cluster ordering. This ordering can be used to inspect the structure and time complexity of cluster. Run time of DBSCAN and OPTICS is nearly same [10].

5. RELATED WORK

Kriegel H.P, Easter M, Sander J. [9] developed the DBSCAN method which was a density based method but not grid based. The basic idea was that for each point in a cluster the neighborhood of radius has to contain at least minimum number of point's .The radius and minimum number are used as input.

Keim D, Hinneburg A [12] developed Denclue a density based algorithm which uses grid. It store only cell information so it is very efficient and contain actual data points. It maintains all the cells into tree based structure. Various extension of the algorithm uses more number of input parameter.

Gehrke J, Agrawal R, Gunopulos [13] developed a technique CLIQUE for high dimensional spaces which is used for data mining. The global density threshold and size of grid is used as input parameter in the clique. This approach find out subspaces which are highest dimensionality and these subspaces contain high density clusters. Huang Z [14] extends the k-means algorithm for categorical domain and mixed attribute domain. Huang purposed two algorithms k-prototype and k-modes. A new distance measure based on total mismatches is used in

categorical domain for k-modes algorithm. A weighted sum of Euclidean distance is used in k-prototypes.

Kumar V, Karypis G, Han Eh [15] merge the Chameleon features to find the almost similar pair of clusters. It combines the interconnectivity and closeness .To find the cluster of data set two phase algorithm is always used. In first phase data set are partitioned into small but relative sub clusters by using partitioning algorithm. In the second phase initial clusters are made by repeatedly joining the small sub clusters by using some algorithm.

Fang D, J Chen et. al. [16] introduced their clustering algorithm which is accessible commercially in data mining tool. The probabilistic model is used for deriving the distance measure. Birch is the base of their clustering algorithms. Birch work well for clusters which are spherical. Xiaofeng H. and Chris D. [17] propose how to select next cluster for merging and splitting in clustering and give several new methods for selection of cluster at next step.

Ting Zhao, George K. [19] present constrained agglomerative algorithms that reduce the errors in agglomerative algorithm hence improve the quality of clustering. Sung H. Myaeng, Bashar AI Shboul [21] propose two algorithms to solve the initialization problem in k-means. Brain Eriksson, Aarti Singh et. al. [22] give an active clustering method which is robust for a controlled fraction of anomalous similarities.

Zahra A V, MarjanKuchaki R et. al. [25] gives an overview of some specific hierarchical clustering algorithms. They classified the algorithms but the main concern was on hierarchical clustering algorithms. They have stated about disadvantage, advantage and attributes etc. of algorithms. The main goal of telling about these characteristics was to minimize the disk input output operation and reducing complexity.

Dibya Jyoti B., Dr. Anil K.G. [24] give an overview of k-means clustering algorithm and different distance measures method. They experimentally show which method is best for their chosen data type. Shede and Bide [26] proposed a clustering pipeline to better the performance of k-means clustering. The originator used a divide and conquers approach for clustering the documents in a data set of 20 newsgroups. Firstly documents were split into groups then a preprocessing and feature extraction take place on each group. K-means clustering applied and cosine similarity measure were used to find out the similarity in documents.

Authors	**Area of contribution**
Fang D, Chen J et. al.(2001)[16]	They developed clustering algorithm for large database which are scalable in nature and work for variant type of attributes.
Xiaofeng He ,Chris ding (2002)[17]	They proposed several new methods for selecting the cluster, which is chosen for merging or splitting in hierarchical clustering. More work can be done on maintaining the cluster balance, because it Improves the performance of clustering.
S.Staab,G.Stumme,A.H.(2003)[18]	Presented an approach in which background knowledge is integrated with the procedure of clustering text document. Background knowledge can be further used in different ways to improve clustering results.
Gerrge Karypis,Ying Z.(2005)[19]	They present a new class of algorithm constrained agglomerative algorithm which merge the good features of partitional and agglomerative approach and it improves the quality of cluster. Some changes can be made into constrained agglomerative methods so that they can perform better than partitional methods.
Ali Ridho B., Kohei Arai (2007)[20]	They proposed a hierarchical k-mean algorithm which improves the starting centroids for k-means by utilizing all the clustering results of k-means in satisfied time. This algorithm works in a better way only for the cases in which it is difficult to make clusters.
Sung H. Myaeng, Bashar AI Shboul (2009) [21]	To solve initialization problem in k-mean they propose two algorithms Initializes Genetic Algorithm (KIGA) and Genetic Algorithm Initializes KM (GAIK). KIGA is only compared with some limited algorithms and said that it is better but it can also be compared with other algorithms.
Brian E. Gautam D. et. al. (2011) [22]	Give an approach in which hierarchical clustering can be done in few number of similarities if the intercluster similarities is less than the intracluster similarities. Active clustering method is robust but can be used only for bounded irregular similarity. Further research can be done in these methods.
Baridam , Barilee, B. (2012)[23]	They represent a method for checking the usefulness of k-means on biological data. They introduce preprocessor schemes which automatically initialize a reasonable value of k to k-mean algorithm. Only applies for biological data, can be further explored so that applicable to all data types.
Dr. Anil K.G.,Dibya Jyoit B. (2014)[24]	They give the idea that a good method for selecting the distance measure depends on the type of data on which work takes place. Further study can be expended for finding the better distance measure in another partition clustering algorithms like K-Medoids, CLARA etc.

6. CONCLUSION

In this paper various clustering techniques has been discussed. Clustering play an important role in data mining. The overview of various clustering techniques has presented in elaborate manner. Various merits and demerits of main clustering techniques have also been discussed. Since clustering is used in various field so it is important to find out which clustering

technique is more effective. This review provides compressive study of various clustering techniques, which may be used for further research. Like by grouping various historical data future happening can be predicted.

7. REFERENCES

[1] Pavel Berkhin (2000), *Survey of Clustering Data Mining techniques, Accrue* Software, Inc.

[2] Sasirekha, K., and P. Baby. "Agglomerative Hierarchical ClusteringAlgorithm-A."InternationalJournal ofScientific andResearch Publications: 83.

[3] Deepa, M. Sathya, and N. Sujatha. "Comparative Studies of Various Clustering Techniques and Its Characteristics." *Int. J. Advanced Networking and Applications* 5.6 (2014): 2104-2116.

[4] Jiawei Han and Michheline Kamber, Data mining concepts and techniques-a reference book, pg. no.-383-422.

[5] Xu Rui and Donald Vrinshc. "Survey of clustering Algorithms." *IEEE Neural Networks on Tronskshns* 16.3 (2005): 645-67

[6] Elavarasi, S. Anitha, J. Akilandeswari, and B. Sathiyabhama. "A survey on partitionclustering algorithms." *International Journal of Enterprise Computing and Business Systems* 1.1 (2011).

[7] Jain, Anoop Kumar, and Satyam Maheswari. "Survey of recent clustering techniques in data mining." *Int J Comput Sci Manag Res* 3 (2012): 72-78.

[8] Lior Rokach & Oded Maimon, .CLUSTERINGMETHODS

[9] Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." *Kdd*. Vol. 96. No. 34. 1996.

[10] Ankerst, Mihael, et al. "OPTICS: ordering points to identify the clustering structure." *ACM Sigmod Record*. Vol. 28. No. 2. ACM, 1999.

[11] Al-Anazi, Sumayia, Hind AlMahmoud, and Isra Al-Turaiki. "Finding Similar Documents UsingDifferentClustering Techniques." *ProcediaComputer Science* 82 (2016): 28-34.

[12] Hinneburg A., Keim D.: *"An Efficient Approach to Clustering in Large Multimedia Databases with Noise"*, Proc. 4th Int. Conf. on Knowledge Discovery & Data Mining, New York City, NY, 1998.

[13] Agrawal R., Gehrke J., Gunopulos D., RaghavanP.: "*Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications*", Proc. ACM SIGMOD'98 Int.Conf. on Management of Data, Seattle, WA, 1998, pp. 94-105

[14] Huang, Zhexue. "Extensions to the k-means algorithm for clustering large data sets with categorical values." *Data mining and knowledge discovery* 2.3 (1998): 283-304.

[15] Karypis, George, Eui-Hong Han, and Vipin Kumar. "Chameleon: Hierarchical clustering using dynamic modeling." *Computer* 32.8 (1999): 68-75.

[16] Chiu T, Fang D, Chen J, Wang Y, Jeris C. A robust and scalable clustering algorithm for mixed type attributes in large database environment. In: Proc 2001 Int Conf on Know-ledge Discovery and Data Mining (KDD'01), SanFrancisco, CA; 2001. pp 263–268.

[17] Chris ding and Xiaofeng He (2002), *Cluster Merging And Splitting In Hierarchical Clustering Algorithms.*

[18] A. Hotho, S. Staab, and G. Stumme. Wordnet improves text document clustering. In Proceedings of the SIGIR Semantic Web Workshop, Toronto, 2003.

[19] Zhao, Ying, George Karypis, and Usama Fayyad."Hierarchical clustering algorithms for document datasets." *Data mining and knowledge discovery* 10.2 (2005): 141-168.

[20] Arai, Kohei, and Ali Ridho Barakbah. "Hierarchical K-means: an algorithm for centroids initialization for K-means." *Reports of the Faculty of Science and Engineering* 36.1 (2007): 25-31.

[21] Al-Shboul, Bashar, and Sung-Hyon Myaeng."Initializing k-means using geneticalgorithms."World Academy of Science,*Engineering and Technology* 54.30 (2009): 114-118.

[22] Eriksson, Brian, et al. "Active Clustering: Robust and Efficient Hierarchical Clustering using Adaptively Selected Similarities." *AISTATS*. Vol. 8. 2011.

[23] Baridam B, Barilee. More work on K -Means clustering algorithm: The dimensionality problem. International Journal of Computer Applications. 2012; 44(2): 23–30.

[24] Bora, Mr, et al. "Effect of different distancemeasures on the performance of K-means algorithm: an experimental study in Matlab." *arXiv preprint arXiv:1405.7471* (2014).

[25] MarjanKuchaki Rafsanjani, Zahra Asghari Varzaneh, Nasibeh Emami Chukanlo (2012), *A survey of hierarchical clustering algorithms, The Journal of Mathematics and Computer Science, 5,.3,* pp.229- 240.

[26] Bide, P., Shedge, R. Improved Document Clustering using k-means algorithm. In: *2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*. 2015, p. 1–5.