# Text Clustering Algorithms: A Review

Himanshu Suyal
Computer Science and Engineering Department,
GBPEC, Pauri garhwal

Amit Panwar
Computer Science and Engineering Department,
GBPEC, Pauri garhwal

Ajit Singh Negi
Computer Science and Engineering Department,
Graphic Era University,
Dehradun

## ABSTRACT

With the growth of Internet, large amount of text data is increasing, which are created by different media like social networking sites, web, and other informatics sources, etc. This data is in unstructured format which makes it tedious to analyze it, so we need methods and algorithms which can be used with various types of text formats. Clustering is an important part of the data mining. Clustering is the process of dividing the large &similar type of text into the same class. Clustering is widely used in many applications like medical, biology, signal processing, etc. This paper briefly covers the various kinds of text clustering algorithm, present scenario of the text clustering algorithm, analysis and comparison of various aspects which contain sensitivity, stability. Algorithm contains traditional clustering like hierarchal clustering, density based clustering and self-organized map clustering.

## General Terms

Text clustering, supervised and unsupervised clustering

## Keywords

Data mining, K mean clustering, text cluster, Hierarchal clustering, prototype, Density bases clustering

## 1. INTRODUCTION

With the large use of computer and rapid development of Internet technology, a large amount of unstructured data is generated by various devices & application. This data is stored in various systems. This system has a lot of information and to extract the appropriate information from this system manually is an impossible task for any organization. Clustering can be very useful to remove the above problem. Text clustering is a process to divide the text content into different clusters according to their similarity like cosine similarity, dice similarity so text clustering is to find that which document has most common words in which document. In order to extract useful information, clustering has become the hot topic for research [1]. Clustering also useful for short text classification which is increasing as the growth of internet [2] Most clustering algorithm used the vector space model (VSM) [3], where text D is consider as a vector in VSM and have a high dimension so choice of the cluster point is very difficult. Some of the text clustering algorithm uses the frequent data item [4].Clustering can be used for various number of task.

**Organization of Document:** For systematically browsing or searching of data hierarchical organization of document into relevant category is very useful. Classic examples of this is used in scatter/ gather method in which a systematic document browsing is done by using document clustering technique [5].

**Corpus summarization:** Clustering technique can be used for corpus summarization by providing coherent summary of the collection in the form of word cluster [6] [7], which can be used to provide summary of the entire content of the underlying corpus.

**Document Classification:** Clustering is an unsupervised learning method where there is no need of training; it can be used to enhance the quality of the supervised variant. Word cluster can be used to improve the performance of the supervised application using clustering technique.

Rest of the paper is organized like this, section II contains representation of text, section III contains different text clustering algorithm, section IV contains performance comparison of text clustering algorithms and section V contains the conclusion.

## 2. TEXT REPERSENTION

Most of the time, text information contains limited structure even no structure, because document contains humans natural language, thus the question that arises in the text clustering is how text representations can be structured so that it can be processed easily and can be analyzed mathematically. A text document can be represented either by binary data where a binary vector maintains the presence or absence of text in the document.

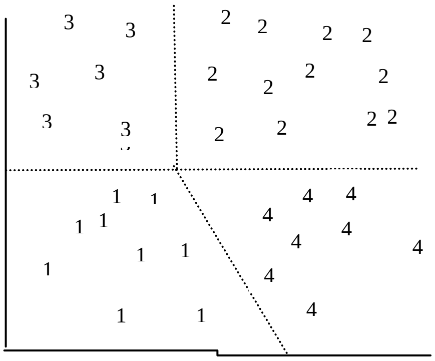

**Figure 1: description of clustering rule**

In this case, text clustering algorithm [8, 11, 13] can be directly used to cluster the data on binary representation. More efficient way to represent the text data that it includes refined weighting method based on the frequency of the individual word in the document as well as it contains the frequency of the word in the entire collection of text (e.g., Term frequency (TF) weighting [9], TF-IDF (Term frequency inverses document frequency) weighting[12]). Most and widely used method is V S M [3] proposed by Professor

Salton. Initially it contains the four clusters. The main idea is to represent the document as point in space (Vector in a vector space), where Space can be represented as collection of the points, points which are similar to each other are classified together and points which are not similar to each other are apart in the space. Figure 1 shows the kind of data division generated by the decision tree algorithm.

# 3. CLUSTERING ALGORITHM

## 3.1 Hierarchal clustering

Text clustering is a kind of typical unsupervised learning. Hierarchal clustering [12] is most common method of clustering which aims building of hierarchy of cluster. The basic concept of hierarchal clustering is to successively merge each document into the predefined cluster based on their similarity. Similarity can be of many types like cosine similarity, dice similarity, etc. The result of hierarchal clustering is to create a cluster hierarchy or dendogram in which leaf nodes are corresponding to the individual document and internal nodes corresponding to the merged group of cluster. When merging of two group is done, a new node is created which corresponds to the larger merged group and children corresponds to the two merged groups.

Hierarchal clustering can be of two types, bottom up and top down. Bottom up hierarchal clustering uses the integration method and called agglomerative clustering, whereas on the other hand top down clustering uses the splitting method and called divisive clustering. The main advantage of hierarchical clustering is the high accuracy but when each class is merged, it is compared to all the classes and the most similar two classes are selected, which makes it relatively slow. The main disadvantage of this clustering is that after doing the merge or split operation it can't be revoked, so it can't correct the wrong decision.

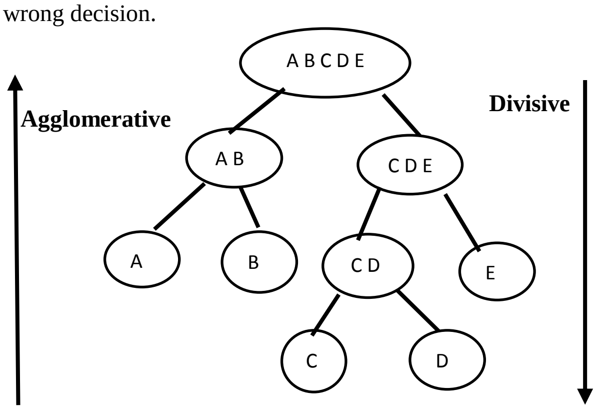

Figure 2: A simple hierarchal cluster

### 3.1.1 Bottom up hierarchical clustering (Agglomerative)

Bottom up hierarchal clustering is based upon the merging technology in which algorithm starts with a single object as a separate category and repeatedly merges the two appropriate categories until it will not meet some stopping condition. Usually number of parameters is k so whenever k category obtains, algorithm doesn't loop. Bottom up hierarchal clustering can be seen as a process of constructing a tree which contains the hierarchy information of the class, and the similarity among all the classes. For a document sets D = {$d_1$, $d_2$, $d_3$, $d_4$, . . .$d_n$}, the agglomerative clustering process is following :

(1) For each document $d_i$ in document set D consider a single member of the class ci = {$d_i$} from cluster C= {$C_1$,$C_2$,$C_3$,…$C_n$}

(2) Calculate the similarity for each classes ($C_i$,$C_j$)in C, similarity can be denoted by sim($C_i$,$C_j$). Similarity can be achieved by cosine similarity, Euclidian distance, Manhattan distance, maximum distance, etc.

(3) Select the largest class of similarity on ($C_i$, $C_j$) into a new class $C_m$so $C_m$= {$C_i$, $C_j$}so now cluster set become C= {$C_1$,$C_2$,$C_3$,…$C_{n-1}$}.

(4) Repeat the above steps until cluster C has only one class.

Advantages of agglomerative clustering is that it can be applied on any shape and it can use any form of similarity, and the main disadvantage of algorithm is that in general case agglomerative clustering have a complexity of almost $\Theta(n^3)$ which causes slowing down of big data set as well as determination of the termination condition of the algorithm making it a tedious task which needs some human expertise.

### 3.1.2 Top- down hierarchical clustering (Divisive)

Top down clustering starts with the one cluster and splitting take place recursively as one down the hierarchy. There are two ways to divide the cluster. One way is optimal solution in which ,first all object belongs to same class $C_1^{(0)}$ =C that can be directly divided into k cluster another way is that $C_1^{(0)}$ can be divided into two classes say $C_1^{(1)}$ , $C_1^{(2)}$ and $C_1^{(1)}$ , $C_1^{(2)}$ can further be divided into more class until class will not divide into k cluster. Divisive clustering have a complexity of almost $\Theta(2^n)$ which make it more worst in comparison to the agglomerative clustering .

Hierarchical clustering is widely used due to its simplicity, flexibility and has an advantage to use any kind of similarity measurements. The most disadvantage of this algorithm is that it is difficult to determine the termination condition.

## 3.2 Distance based partitioned clustering

Distance based partitioned clustering is widely used in database literature for creating cluster of object efficiently. The objective of partitioned clustering is to divide the data set into k disjoint cluster by using the distance measurements. Each k cluster contains the homogeneous data. Homogeneity can be achieved by using similarity. Partitioned clustering is also known as the k-estimate or c-estimate due to the tendency to divide the data set into fixed k or c cluster. It will not give guarantee to local and global optimal solution because number of data point in any data set is always finite and number of distinct cluster is finite, the problem of local minima can be removed by using exhaustive search. The most widely used distance based partitioned clustering algorithm is k-mediod [10] &k-mean [15].

### 3.2.1 k-mean algorithm

The goal of k-means algorithm is based on the input parameters k, in which the data set is divided into k clusters. Algorithm uses iterative update in each round, based on k point of reference points which were grouped around k clusters. Each cluster centroid will be used as a reference point for next round of iteration. Iteration makes the selected reference point closer to the true cluster centroid, so the clustering effect gets better.

K-mean algorithm is as follows:

Suppose a set of data point D= $\{x_1, x_2, x_3, \ldots x_n\}$,$x_{i=}$ $\{x_{i1}, x_{i2}, x_{i3 \ldots} x_{im}\}$ is vector in real space $X \subseteq \Re$ where m represents the number of the data.
Input: D documents to be clustered, the cluster number k
Output: k clusters, and each document will be assigned to one cluster
Algorithm of k-mean (k, D)

1) Firstly choose the k data point as an initial centroid.

2) Repeat for each data point when $x \in D$.

3) Compute the distance of x to each centroid and assign x into closest centroid.

4) Repeat the step until it doesn't meet some stop condition.

k-mean algorithm have advantages as the following : simple and uses small number of iteration even less than 5 iteration is sufficient for large data set; can run parallel ;good effect on the convex cluster. But the main disadvantage of k-mean clustering is that it's very much dependent on the initial choice of the cluster. Other disadvantages are as follows: sensitive to the isolated point; can't discover the non-sphere cluster; very frequently falls on the optimum solution; less stability; sometimes clustering remains unbalanced. To overcome this problem, a new improved k-mean algorithm came into picture [15].

### 3.2.2 K-mediod algorithm

K-mediod algorithm used the set of data point from the original data as the anchor or mediod around which cluster is built. The main aim of the clustering algorithm is to determine the optimal solution from the original data set around which the cluster is built. Process of k-mediod algorithm is similar to the k-mean but main difference between two algorithms is that k-mean algorithm uses centroid to represent the cluster and k-mediod represents the cluster using object closest to the center. K-mediod uses the iterative approach in which the use of randomized inter-changes k representative are successfully improved .It uses the average similarity of each document represented as the objective function which needs to be improved during the interchange process. In each iteration, it replace a randomly picked representative in the current set of mediod with a randomly representative from the collection. It improves the clustering objective then it is applied to the iteration until the convergence. The main disadvantage of k-mediod algorithm is that it requires large number of iteration in comparison to the k-mean algorithm .This is because in each iteration, it requires the computation of the objective function other disadvantage of the k-mediod it will not work well on the sparse data, this is because sparse data have less similarity.

## 3.3 Density Based Clustering

It is very difficult to cluster when data is in nonlinear shape means in different size, density and shape, so k-mean and k-mediod algorithm can't be applied on these type of data. Density based clustering algorithm is very useful to classify these type of data, in density based clustering algorithm cluster is defined as area of higher density that is the remainder of the dataset. Density based clustering mainly considers density and boundary area of the cluster .The most widely used density based spatial clustering with the application of noise is DBSCAN and DENCLUE algorithm.

### 3.3.1 DBSCAN Algorithm

The DBSCAN algorithm [12] is density based clustering algorithm which uses density function and widely used for the cluster of arbitrary shape. DBSCAN exploits the fact that cluster is a group of objects which are density reachable from the arbitrary core of object in the cluster. Density based object can be retrieved by literarily collecting directly density reachable object. DBSCAN checks in database for the each point of ε–neighborhood. If ε–neighborhood Nε(o) of a point has more than μ elements. Where o is called core point, and an object Nε(o) is created which is hold by a new cluster c, then the ε–neighborhood of all point p in c is checked which have not yet to be processed . If Nε(o) contains more than μ points, then neighbor of p which are not already contained in c are added to the cluster and their ε–neighborhood is checked in the next step this procedure is repeated until no new point can be added to the current cluster c. The advantage of the DBSCAN is that no need to define number of cluster in advance, it can be used to cluster arbitrary shape cluster even it can be cluster which is surrounded by different cluster. The main disadvantage of this is that it can't cluster data with large density.

### 3.3.2 DENCLUE Algorithm

DENCLUE algorithm [17] depends on a function called by influence function. Influence function is a mathematical function between each data point and other adjacent data points; it is used to quantify a value in their areas. Influence function is modeled by the density of an object at the data space of all data objects at the object space. This function uses the density attraction points to obtain clustering resulting from the clusters by using divide and merge operation. Global density function has a maximum locals called as density attraction point. DENCLUE has several advantages over other algorithms. It is based on strong mathematical concept; it provides good clustering results for large and noisy data sets; it uses simple mathematical description for clusters of arbitrary shape; it work on the principle of unit organization data to efficiently handle large high dimensional data.

## 3.4 Organizing Maps Algorithm

Self- organizing maps algorithm is used to simulate the characteristics of the human brain to the signal processing and can be used to develop an artificial neural network. Finnish Helsinki professor Teuvo Kohonen in 1981 proposed this model, and nowadays it become famous and widely used in the self-organizing neural network. Self – organizing maps (SOM) is also known as Kohenen network. The SOM [16] algorithm as follow:
(1) Randomly choose initial connection weights and set the maximum times for the K
(2) Initialize the training counter k==0
(3) Randomly select input mode and calculate the Euclidian distance for each input unit.
(4) Select Get node.
(5) Makes some adjustment onto the connection weight of wining node and its domain.

(6) Counter k is incremented, if k<K go to the step 3 otherwise end train.

(7) Output the result.

## 4. COMPARISION

Clustering is the process to divide the data sets into the similar group. The performance of clustering algorithm depends on the following criteria.

1) It should be highly flexible; it produces good results for small data sets as well as for large data sets.

2) It should handle the high dimensional data very well. Text data have very high dimensions so clustering algorithm produce good result on the high dimensional data.

3) It should behave well on noisy data. Large majority of the data base contain noisy data so clustering algorithm should not degrade the quality of the clustering result.

4) It should not influence by the initial choice of cluster as well as the domain of the data.

5) It should have less complexity.

Table 1 showing the comparison result of different clustering algorithm based on some parameter. Table 2 showing the complexity of the discussed algorithm

**Table 1: comparison of different clustering algorithm**

| Criteria | Hierarchal clustering | K-mean | K-mediod | DBSCAN | DENCLUE | Self- organizing map |
|---|---|---|---|---|---|---|
| Initial condition | No | Yes | Yes | Yes | Yes | Yes |
| Termination condition | Not precise | Precise | Precise | Precise | Precise | Precise |
| Arbitrary value | No requirement | Numeric attribute | Numeric attribute | Numeric attribute | Numeric attribute | - |
| Effect on Size of data sets | Not good | Good | Not good | Not good | Not good | Good |
| Shape of data set | Arbitrary | Convex | Convex | Arbitrary | Arbitrary | - |
| Granularity | Flexible | K and initial point | K and initial point | Threshold | Threshold | Parameter |
| Result optimization | Optimization | Rebuild optimization | Rebuild optimization | Rebuild optimization | Optimization | Optimization |
| Handling dynamic data | No | Yes | Yes | Yes | Yes | Yes |
| Behavior on noisy data | No influences | Influences | Influences | Not much influences | Not much influences | - |
| Distance measurement | Any | Distance at normal space | Distance at normal space | Density | Density | Euclidian distance |
| Implementation | Simple | Simple | Complicated | Simple | Simple | Simple |

**Table 2: computation complexity of algorithm**

| Algorithm | Complexity |
|---|---|
| Hierarchal clustering | $O(n^3)$ for the agglomerative clustering<br>$O(n^2)$ for the divisive clustering<br>Where n is the number of total object |
| k-mean | O(n k l)<br>Where n =total no number of abject ,<br>k= the number of cluster<br>l=is the number of iteration |

| | |
|---|---|
| **k-mediod** | O( l k (n-k)$^2$)<br>Where n =total no number of abject ,<br>k= the number of cluster<br>l=is the number of iteration |
| **DBSCAN** | O(n log n )<br>Where n= number of the object |
| **DENCLUE** | O(n$^2$ )<br>Where n= number of the object |
| **Self-organizing Map** | O(n k m )<br>Where n= total number of object<br>k= number of neuron<br>m= number of training count |

## 5. CONCLUSION

Clustering is the process to divide a data sets into the predetermine class or cluster, it is widely used in the field of information technology to increases the process of information retrieval as well as it is used in field of natural language processing. This paper introduces the comparison between different text clustering algorithm based on some parameters which can be used to develop the new clustering algorithm which can efficiently cluster the data well as can work on the arbitrary data very well.

## 6. REFERENCES

[1] Yu Hui Document cluestring based on Modified Latent Semantic analysis[j].Journal of chinese Computer System, 2009, 30(5):963-966

[2] Himanshu Suyal and R B Patel. Article: Improved Information Filtering and Feature Dimensionality Reduction using Semantic based Feature Dataset for Text Classification: In Context to Social Network. International Journal of Computer Applications 94(18):42-46, May 2014. Published by Foundation of Computer Science, New York, USA.

[3] Salton G, Wong A, Yang C.A vector space model for automatic indexing[J] .Communications of the ACM, 1975, 18( 11) : 613- 620.

[4] S.Murali Krishna, S.Durga Bhavani. An Efficient Approach for Text Clustering Based on Frequent Itemsets. [J]European Journal of Scientific Research. Vol.42 No.3 (2010), pp.385-396

[5] D. Cutting, D. Karger, J. Pedersen, J. Tukey. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. ACM SIGIR Conference, 1992.

[6] H. Schutze, C. Silverstein. Projections for Efficient Document Clustering, ACM SIGIR Conference, 1997.

[7] R. Bekkerman, R. El-Yaniv, Y. Winter, N. Tishby. On Feature Distributional Clustering for Text Categorization. ACM SIGIR Conference, 2001.

[8] D. Gibson, J. Kleinberg, P. Raghavan. Clustering Categorical Data: An Approach Based on Dynamical Systems, VLDB Conference, 1998.

[9] G. Salton, C. Buckley. Term Weighting Approaches in Automatic Text Retrieval, Information Processing and Management, 24(5), pp.513–523, 1988.

[10] Park, Hae-Sang, and Chi-Hyuck Jun. "A simple and fast algorithm for K-medoids clustering." Expert Systems with Applications 36.2 (2009): 3336-3341.

[11] P. Andritsos, P. Tsaparas, R. Miller, K. Sevcik. LIMBO: Scalable Clustering of Categorical Data. EDBT Conference, 2004.

[12] Ying Zhao; George Karypis; Usama Fayyad. Hierarchical Clustering Algorithms for Document Datasets. Data Mining and Knowledge Discovery [J]. Vol.10, 2005. pp:141-168.

[13] D. Gibson, J. Kleinberg, P. Raghavan. Clustering Categorical Data: An Approach Based on Dynamical Systems, VLDB Conference, 1998.

[14] Easter M.,kriegel H.-P.,sander j.,Xu.: A Density Based Algorithm for Discovering Cluster in Large Spatial data based with noise,KDD'96,pp.226-231.

[15] Xiaojun Wang, Jianwu Yang, Xiaoou Chen. An Improved K-means Document Clustering Algorithm [J] Computer Engineering, 2003, 29(2): 102-104.

[16] Yang Zhanhua,Yang Yan. "Document clustering method based on hybrid of SOM and K_means". Journal of Computer application research, 2008, Vol. 18, No. 8, pp.73-79.

[17] Hinneburg, Alexander, and Daniel A. Keim. "An efficient approach to clustering in large multimedia databases with noise." *KDD*. Vol. 98. 1998.