Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 10, October 2014, pg.58 – 61

RESEARCH ARTICLE

Comparison of Basic Clustering Algorithms

Darshan Sonagara[1], Soham Badheka[2]
[1]Student, G H Patel College of Engineering & Technology, Gujarat, India
[2]Student, Chandubhai S. Patel Institute of Technology, Charusat, Gujarat, India
[1] darshan.sonagara@gmail.com; [2] sohambadheka008@gmail.com

Abstract— This paper presents the results of the theoretical study of some common document clustering techniques. Clustering is a machine learning technique for data mining which is a grouping of similar data for analysis purpose in simple words. We have compared the two main approaches of document clustering that are hierarchical clustering and Partitional clustering algorithm. We have surveyed and listed the algorithms, its advantages and disadvantages as well. Hierarchical clustering and its two basic approaches are discussed which are Agglomerative and Divisive. In partitional clustering, various partitions are generated by the partitioning algorithms like K-Means. However K-Means algorithm is very different from the hierarchical algorithms. Both of the approaches are better depending on the different situations. Partitional clustering is faster than the hierarchical clustering and partitional clustering is based on the stronger assumptions. In contradiction, hierarchical algorithm needs only a similarity measure and does not require input to be given.

Keywords— Document clustering, Clustering algorithms, K-means algorithm, Hierarchical algorithm, Partitional algorithm

I. INTRODUCTION

The goal of the survey is to provide a review of two main clustering techniques in data mining. As the data on the web increases it becomes harder to store them in a meaningful way or to extract some useful information from them so that we need Document Clustering. This large amount of data can be both structured and unstructured which needs to be processed and analyzed. Document clustering is the traditional data mining technique which groups the related documents and organizes them. Today it has become very necessary to apply these techniques on World Wide Web to give a user better experience and a better organization for business analysts.

Generally, there are two very basic clustering models. The first one is the connectivity based model which includes hierarchical based algorithm and another is centroid based model which includes K-Means algorithm.

In the very first section, we are going to mention the classification of the clustering techniques in brief and then we will discuss the algorithms. Moreover we will compare the algorithms and find the most suitable algorithm accordingly.

II. CLUSTERING ALGORITHMS

As mentioned above there are two basic algorithms:

1. Hierarchical clustering
2. Partitional Clustering

1. Hierarchical algorithm

Hierarchical clustering builds a cluster hierarchy or we can explain it by a tree of clusters which is widely known as dendogram. Every node of the cluster or you can say a document on web contains child clusters. There are two basic approaches in hierarchical algorithm.[1]

A) Agglomerative: Begin with the points as individual clusters and at every step, merge the most similar or nearest pair of clusters. This needs a definition of cluster similarity or distance. We will conclude that it is essentially a bottom up approach of hierarchical clustering.

B) Divisive: Begin with one, all-embracing cluster and at every step, split a cluster till solely singleton clusters of individual points stay. In this case, we would like to determine, at every step, which cluster to separate and how to perform the split. Thus essentially it is a top down approach of hierarchical clustering.[2]

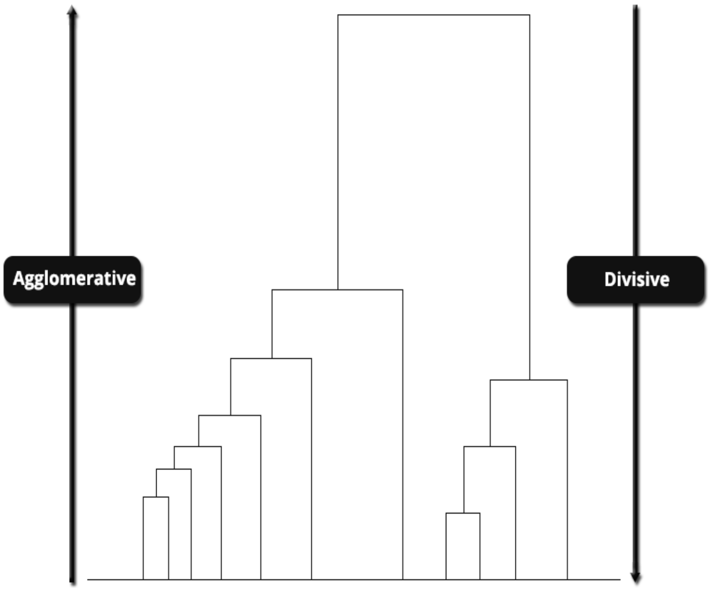

Figure 1. Hierarchical Algorithm

2. Partitional Algorithm:

The other sort of clustering algorithm relies on the centroid based model. It is referred to as a partitional technique as well. There are variety of partitional techniques, however here we have taken K-Means algorithm that is widely used in the context of document clustering. It is a centroid based algorithm stating that a cluster can be depicted as the center point that is the median point of a group of points. Within the context of document clustering these points are the documents itself.

III. COMPARISON OF CLUSTERING ALGORITHMS

1) Aggloromative Algorithm:

1. Compute the similarity between all the pairs of clusters. i.e., calculate the similarity matrix whose ij^th^ entry in matrix gives the similarity between the i^th^ and j^th^ pair of clusters.
2. Combine the foremost similar two clusters.
3. Update the similarity matrix to replicate the pairwise similarity between the new cluster and the original clusters.
4. Repeat steps 2 and 3 until only a single cluster remains. [3]

Advantages of hierarchical clustering :

- Embedded flexibility concerning the extent of granularity.
- Ease of handling of any types of similarity or distance.
- Consequently, the applicability to any attribute varieties.

Disadvantages of hierarchical clustering :

- Vagueness of termination criteria.
- The undeniable fact that most hierarchical algorithms don't revisit.

2) K-Means algorithm:

K-Means is a straightforward learning algorithm among the clustering algorithms. Basically, K-Means algorithm searches out the most effective division of n entities in k groups, in order to find the total distance of the group's members to its corresponding centroid, representative of the cluster, is reduced. Formally, the goal of the algorithm is to partition the n entities into k sets Si where, i=1, 2… k so that the within-cluster sum of squares (WCSS) is minimized, defined as

$$\sum_{j=1}^{k}\sum_{i=1}^{n} ||X_i^j \ - \text{Cj}\,||^2 \qquad (1)$$

Where, term $||X_i^j \ - \text{Cj}\,||$ provides the distance between an entity point and also the cluster's centroid.

Basic K-Means Algorithm for finding *K* clusters.

1. Select *K* points as the initial centroids.

2. Assign all points to the closest centroid.

3. Recompute the centroid of each cluster.

4. Repeat steps 2 and 3 until the centroids don’t change.

The most common algorithm which is described below, uses an iterative refinement approach. The steps are as following:

- Define the initial groups' centroids. A very common way is to assign the random values for the centroids of the groups. Another approach uses the values of *K* different entities as the centroids.
- Assign each entity to the cluster which has the closest centroid. To find the cluster with the most similar centroid, the algorithm must calculate the distance between all the entities and each of the centroids.

- Recalculate the values of the centroids. The values of the centroid's fields are updated, taken as an average of the values of the entities' attributes which are part of the cluster.
- Repeat steps 2 and 3 iteratively until entities do no longer change groups. [4]

Advantages of K-Means Algorithm

- Simple
- Fast for low dimensional data
- It can find pure sub clusters if large number of clusters is specified.
- With a large number of variables, K-Means may be viewed as computationally faster than hierarchical clustering
- K-Means is expected produce tighter clusters than hierarchical clustering

Disadvantages of K-Means Algorithm

- K-Means cannot handle non-globular data of different sizes and densities
- K-Means will not identify outliers
- K-Means is restricted to data which has the notion of a center (centroid)

CONCLUSION

In this survey paper, we presented the theoretical study of some common document clustering techniques. We compared the two basic approaches to document clustering that are agglomerative hierarchical clustering over K-means algorithm and analyzed that the standard K-means approach is better than the hierarchical approaches. Moreover, the K-Means is a centroid based algorithm and it is computationally faster on low dimensional data than hierarchical algorithms. So that it gives tighter clusters and it gives an even distribution to some extent than the hierarchical clustering techniques. Theoretically, the run time of K-means is O(n) compared to hierarchical algorithm which is O(n^2).

Hierarchical algorithms are connectivity based algorithms as they do not allow revisits, while in K-Means algorithm we can specify the number of iterations to get smooth distribution of documents. In contradiction, K-Means algorithm needs the number of clusters to be specified first and for that we need to have a strong assumption. In addition, K-Means algorithm takes random seeds in the first iteration and there might be a situation when the seeds that are chosen diverges instead of converging. But on the whole K-Means gives better clustering if proper assumptions are made and if we could overcome some of the limitations by presenting some refinement to the conventional K-Means algorithm.

REFERENCES

[1] Survey of Clustering Data Mining Techniques by Pavel Berkhin (Accrue Software, Inc.)

[2] Dynamic hierarchical algorithms for document clustering Reynaldo Gil-García , Aurora Pons-Porrata

[3] A Survey of Document Clustering Techniques & Comparison of LDA by Yu Xiao

[4] A Comparison of Document Clustering Techniques Michael Steinbach George KarypisVipin Kumar Department of Computer Science and Engineering, University of Minnesota Technical Report #00-034 {steinbac, karypis, kumar@cs.umn.edu}.

[5] An Efficient K-Means Clustering Algorithm by Khaled Alsabti(Syracuse University), Sanjay Ranka (University of Florida), Vineet Singh(Hitachi America, Ltd.)