Agglomerative Clustering Implementation for Information Retrieval

Introduction:

Locating & fetching the relevant information is key to any Information Retrieval system. In general, Search Engines takes the user query as the input and gives the ranked listed of documents as the output. The ranked list thus sorts the documents in the order it was ranked, letting the user browse through the results based on the ranked list. The ranked list is based on a set of assumptions that users scans through the contents sequentially & the ranked documents are independent.

An alternative approach to organizing & retrieving the documents is based on the Cluster Hypothesis: "closely associated documents tend to be relevant to the same requests". There are different approaches for clustering the documents like Flat Clustering, Hierarchical Clustering, Grid based Clustering, Density based Clustering etc.,. In this review, we will see about one of the Hierarchical Clustering method – Agglomerative Clustering, its implementation, its pros and cons & applications in Information Retrieval.

Body

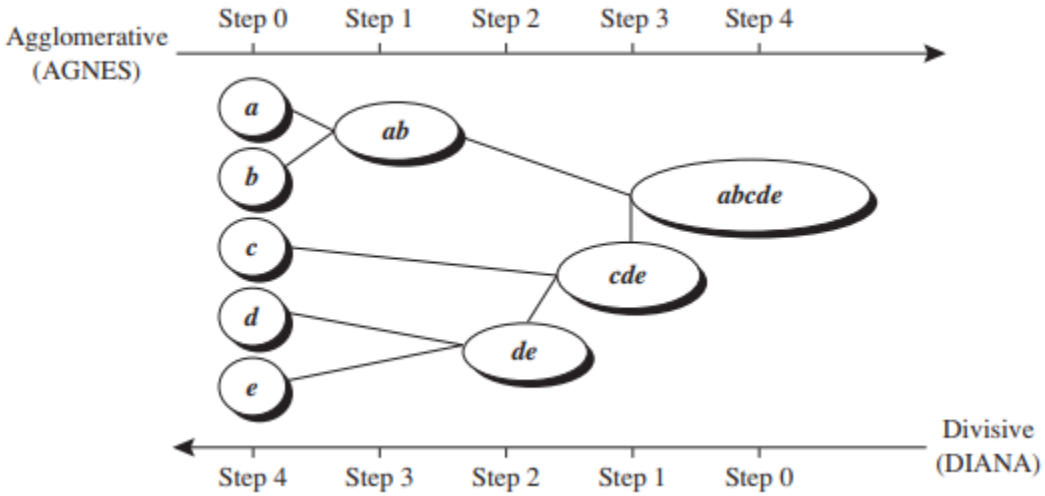
What is Agglomerative Clustering?

Before we answer the question on the Agglomerative Clustering in specific, let us start with a small definition of Hierarchical Clustering. In hierarchical clustering, the data objects into a hierarchy or "tree" of clusters. Representing documents or collections in the form of a hierarchy is useful for summarization and visualization.

Agglomerative Clustering is a hierarchical clustering where the clustering is done bottom up. Initially, each document is placed in a cluster of its own. Then the individual document clusters are grouped based on the similarity of the clusters iteratively to form hierarchy of clusters.

Below figure diagrammatically represents the AGNES (Agglomerative Nesting of Clusters)

Clustering:



Measuring distance between clusters:

There are multiple ways to measure the similarity or the dissimilarity between the clusters.

'Euclidian' measure is a commonly used distance measure. Other similarity measures that can be used are Cosine Similarity, Manhattan distance etc.

Linking Clusters:

When the different clusters are merged iteratively, there are several approaches that can be taken how the clusters are linked.

Single Linkage: The distance between two clusters is defined as the minimum value of all pairwise distances between the elements in two clusters. It tend to produce loose clusters.

Complete Linkage: The distance between two clusters is defined as the maximum value of all pairwise distances between the elements in two clusters. It tends to produce more compact clusters, as the clusters are linked based on the object in the cluster that is far apart.

Average Linkage: The distance between two clusters is defined as the average distance between the elements in the two clusters.

Centroid Linkage: The distance between two clusters is defined as the distance between the centroid of the two clusters

Implementation of Document Clustering using Agglomerative Clustering:

To implement document clustering using Agglomerative Clustering algorithm:

1. Extract the text from the documents – Extract all distinct text from the set of documents to be clustered
2. Preprocess or clean the data – This would include stemming & removal of the stop words from text extracted
3. Build a Vector Space Model - To represent document as a vector. The Term Frequency & the Inverse Document Frequency can be implemented when building the Vector Space Model
4. Define the Distance Method & the Cluster Linkage method to be used in for Clustering

Pros of Agglomerative Clustering:

1. Don't have to define the number of clusters beforehand
2. Easy to understand and produces a good visualization using dendogram
3. Level of Clustering is flexible

Cons of Agglomerative Clustering:

1. Computationally intensive
2. Does not perform well with missing data or outliers
3. Once a mistake is made in forming the clusters, it cannot be corrected

Conclusion:

In this review we have thus seen the details of the Agglomerative Clustering, the methodologies to be considered while using agglomerative clustering based on the cluster needs, its implementation, Pros and Cons. While Agglomerative clustering has its own advantages, it is computationally intensive & performs poorly with the outliers. Other clustering algorithm like K Means, Improved K Means, Expectation Maximization, and Density based clustering (DB SCAN) gives competitive results and has its own advantages & disadvantages over Agglomerative Clustering. It is better to choose the algorithm for clustering based

on the dataset that is used, with good understanding of each algorithms advantage & disadvantage.

References:

Data mining: Concepts and Techniques - Jiawei Han, Micheline Kamber, Jian Pei

https://towardsdatascience.com/machine-learning-algorithms-part-12-hierarchical-agglomerative-clustering-example-in-python-1e18e0075019

https://www.datanovia.com/en/lessons/agglomerative-hierarchical-clustering/#:~:text=The%20agglomerative%20clustering%20is%20the,as%20AGNES%20(Agglomerative%20Nesting).&text=Next%2C%20pairs%20of%20clusters%20are,big%20cluster%20containing%20all%20objects