Comparison of AGNES and DIANA Clustering Techniques

Introduction

Text classification is a vastly important task in speeding up our searches amongst data which is increasingly becoming massive everyday. Gone are the days when the simple technique of just measuring the relevance of texts to a user query would have been sufficient. Such a task would put a great strain in the computation efforts in a large database like the world wide web. Clustering multiple similar documents and creating one representative of that cluster would vastly reduce the number of comparisons required for finding out relevant documents. Thereby, precision and recall can be improved as well if clustering could be done successfully as all the relevant documents would be close by in the vector space. Clustering can also be used to create a classifier which can classify new documents into the clusters after creating clusters for already present documents in the collection.

There are quite a few clustering methods, two of which are broadly used nowadays: partitional and hierarchical clustering. The most common partitional clustering is k-means. Hierarchical clustering in turn can be implemented in two ways: Agglomerative and Divisive. Agglomerative Clustering uses a "bottom-up" approach whereas Divisive Clustering uses a "top-down" approach.

Agglomerative Clustering (AGNES)

Agglomerative Clustering or AGglomerative NESting (AGNES) starts with all objects as a cluster. Hence there will be N clusters where N is the number of objects and then they will be iteratively merged into clusters according to their similarity distances. Similarity distances can be calculated using many options: Euclidean, Cosine distances to name a few. The result of this clustering technique will be a dendrogram, which is plotted with all the objects in the x-axis and the height in the y axis.

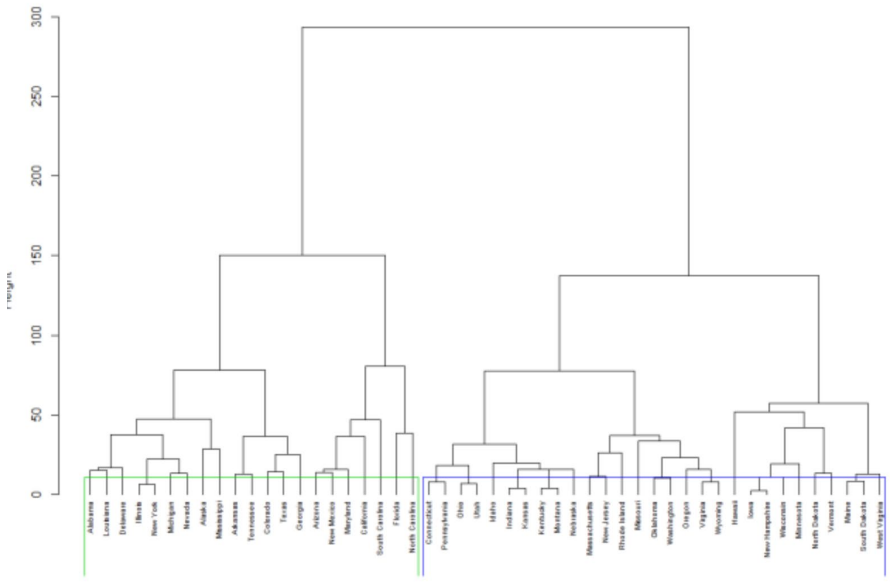

Fig 1: Dendrogram

The height indicates the dissimilarity distance between two clusters. The higher the height, the more dissimilar the clusters are. The cut off point would depend on the number of clusters desired.

Divisive Clustering (DIANA)

Divisive Clustering or DIVisive Analysis is just the opposite of AGNES. Where AGNES uses the bottom up approach, DIANA goes from one big cluster which will have all the objects and recursively divide it to the number of clusters required according to their dissimilarity distance. Needless to say, DIANA is more

capable of finding out large clusters as they start from one big cluster while AGNES is more successful in discovering smaller clusters as it starts from N clusters.

The most important question in this method of creating clusters is how to calculate the dissimilarity. The five most common methods are :

1. Maximum or complete linkage clustering: Considers the largest value of the pairwise dissimilarities as the cluster distance
2. Minimum or single linkage clustering: Considers the smallest value of the pairwise dissimilarities as the cluster distance
3. Mean or average linkage clustering: Considers the average value of the pairwise dissimilarities as the cluster distance
4. Centroid linkage clustering: Considers the distance between centroid values as the cluster distance
5. Ward's minimum variance method: Minimizes the within-cluster variance.

The hierarchical clustering methods are notoriously slow as compared to partitional clustering as it has a time complexity of $O(n^2)$ where n is the size of the Collection which will be huge for text retrieval purposes. However, they tend to give better results than partitioning methods like k-means, albeit this has been tested predominantly in non-document atmospheres.

Hierarchical clustering is extremely computation heavy when the dataset is large. One way of implementing this is to use user defined heuristics along with AGNES or DIANA. Clustering can be done for newly found pages in an offline manner which would classify the pages while it is being crawled. Ranking of the most relevant pages of the cluster can be done and then when the query term is matched with the cluster representative then all the top ranked documents can be displayed to the user. This involves a lot of work in the background by the search engines however, the quality of the information provided can be greatly improved using this approach.

Conclusion

Hierarchical Clustering is not a desired approach if the target is to achieve quick results in real time. There has been numerous approached proposed in various papers where this is implemented along with other methods, e.g. k-means which requires less computation. One interesting proposal is to use "bisecting" k-means clustering which would bisect the largest cluster using k-means and continue doing that until the desired number of clusters are formed. This form is sometimes even more efficient than k-means, especially when the number of clusters is large. All in all, AGNES and DIANA are both very similar clustering methods which could result in high quality information retrieval at the expense of time complexity. However, if utilized with other efficient methods, hierarchical clustering can be very effective.

References:

1. Beil, Florian, Martin Ester, and Xiaowei Xu. "Frequent term-based text clustering." *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2002.
2. https://bradleyboehmke.github.io/HOML/hierarchical.html
3. https://www.datasciencecentral.com/profiles/blogs/usarrests-hierarchical-clustering-using-diana-and-agnes
4. Iwayama, Makoto, and Takenobu Tokunaga. "Hierarchical Bayesian clustering for automatic text classification." Proceedings of the 14th international joint conference on Artificial intelligence-Volume 2. 1995.