Abstract

Document clustering is a traditional, efficient and yet quite effective, text mining technique when we need to get a better insight of the documents of a collection that could be grouped together. The K-Means algorithm and the Hierarchical Agglomerative Clustering (HAC) algorithm are two of the most known and commonly used clustering algorithms; the former due to its low time cost and the latter due to its accuracy. However, even the use of K-Means in text clustering over large-scale collections can lead to unacceptable time costs. In this paper we first address some of the most valuable approaches for document clustering over such 'big data' (large-scale) collections. We then present two very promising alternatives: (a) a variation of an existing K-Means-based fast clustering technique (known as BigKClustering - BKC) so that it can be applied in document clustering, and (b) a hybrid clustering approach based on a customized version of the Buckshot algorithm, which first applies a hierarchical clustering procedure on a sample of the input dataset and then it uses the results as the initial centers for a K-Means based assignment of the rest of the documents, with very few iterations. We also give highly efficient adaptations of the proposed techniques in the MapReduce model which are then
experimentally tested using Apache Hadoop and Spark over a real cluster environment. As it comes out of the experiments, they both lead to acceptable clustering quality as well as to significant time improvements (compared to K-Means - especially the Buckshot-based algorithm), thus constituting very promising alternatives for big document collections.

References

7. Raymond T. Ng and Jiawei Han. CLARANS: A method for clustering objects for spatial data mining. IEEE Transactions on Knowledge and Data Engineering, 14(5):1003–1016, 2002.
8. Raymond T. Ng and Jiawei Han. Efficient and effective clustering methods for spatial data mining. In VLDB, pages 144–155, 1994.

Index Terms
Keywords

Document Clustering, Big Data, KMeans, Hierarchical Clustering, MapReduce Model, Hadoop, Spark