Abstract

Since the amount of text data stored in computer repositories is growing every day, we need more than ever a reliable way to assemble or classify text documents. Clustering can provide a means of introducing some form of organization to the data, which can also serve to highlight significant patterns and trends. Document clustering is used in many fields such as data mining and information retrieval. This paper presents the results of an experimental study of some common document clustering techniques. In particular, we compare the two main approaches of document clustering, agglomerative hierarchical clustering Modified BIRCH and Partitional clustering algorithm K-means. As a result of comparing both algorithms we attempt to establish appropriate clustering technique to generate qualitative clustering of real-world document.

References


Index Terms

Computer Science

Algorithms

Keywords