Abstract

In the field of data mining, the approach of assigning a set of items to one similar class called cluster and the process termed as Clustering. Document clustering is one of the rapidly developing, research area for decades and considered a vital task for text mining due to exceptional expansion of document on cyberspace. It provides the opportunity to organize a large amount of scattered text, in meaningful clusters and laydown the foundation for smooth descriptive browsing and navigation systems. One of the more often useable partitioning algorithm is k-means, which is frequently use for text clustering due to its ability of converging to local optimum even though it is for enormous sparse matrix. Its objective is to make the distance of items or data-points belonging to same class as short as possible. This paper, exploring method of how a partitioned (K-mean) clustering works for text document clustering and particularly to explore one of the basic disadvantage of K-mean, which explain the true value of K. The true K value is understandable mostly while automatically selecting the suited value for k is a tough algorithmic problem. The true K exhibits to us how many cluster should make in our dataset but this K is often ambiguous there is no particular answer for this question.
while many variants for k-means are presented to estimate its value. Beside these variants, range of different probing techniques proposed by multiple researchers to conclude it. The study of this paper will explain how to apply some of these techniques for finding true value of K in a text dataset.

References

47. M. Meila, and D.Hackerman 1998, “An Experimental Comparison of Several Clustering and Initialization Method” Microsoft Research Redmond, WA.

Index Terms

Computer Science

Algorithms

Keywords
K-Means, Clustering, Unsupervised Learning, Pre-processing