Abstract

Extremely large data sets often known as ‘Big Data’ are analyzed for interesting patterns, trends, and associations, especially those relating to human behavior and interactions. Extraction of meaningful and useful information needs to be done in parallel using advanced clustering algorithms. In this paper, effort has been made to tweak changes to the existing K-means algorithm so as to work in parallel using MapReduce paradigm. K-means due to its gradient descent nature is highly sensitive to the initial placement of the cluster centers. This random initialization of cluster centers results in empty clusters and slower convergence. In this paper, an overview of existing methods with emphasis on computational efficiency is presented. Comparison of three well known linear time complexity initialization methods has been presented here. These methods are analyzed on two different data sets. The experimental results are recorded and presented with insights on different initialization methods for practitioners.
17. S. Sakr, A. Liu, and A. G. Fayoumi. “The family of MapReduce and large-scale data
A Comprehensive Survey on Centroid Selection Strategies for Distributed K-means Clustering Algorithm


**Index Terms**

Computer Science

Algorithms

**Keywords**

K-means Clustering Algorithm, Hadoop, MapReduce, PCA, HDFS.