Abstract

In this In today's era data generated by scientific applications and corporate environment has grown rapidly not only in size but also in variety. This data collected is of huge amount and there is a difficulty in collecting and analyzing such big data. Data mining is the technique in which useful information and hidden relationship among data is extracted, but the traditional data mining approaches cannot be directly used for big data due to their inherent complexity. Cluster analysis is used to classify similar objects under same group. It is one of the most important data mining methods. However, it fails to perform well for big data due to huge time complexity. For such scenarios parallelization is a better approach. Map reduce is a popular programming model which enables parallel processing in a distributed environment. In this paper to propose system for analyze the performance of two clustering techniques on big dataset. The goal of this paper is to find better clustering technique between K-Medoid and
BIRCH clustering by applying on real life large dataset.

References

- Hae-Sang Park, Chy Hyuck Jun, A Simple And Fast Algorithm For K-Medoid clustering, Department of Industrial and Managerial Engineering, POSTECH, San 31 Hyojado-dong, Pohang 790-784, South Korea.
Analysis of Clustering Techniques on Big Data

Index Terms

Computer Science
Information Sciences

Keywords
Big Data  Clustering  Hadoop  Map-reduce  K-medoid(km)  Birch.