Chapter 15
Hybrid Clustering Technique to Cluster Big Data in the Hadoop Ecosystem:
Big Data Application

E. Padmalatha
Chaitanya Bharathi Institute of Technology, India

S. Sailekya
Chaitanya Bharathi Institute of Technology, India

ABSTRACT

Big data analytics as well as data mining play vital roles in extracting the hidden statistics. Customary advances for investigation and extraction of hidden information from data may not exert efficiently for big data because of its complex, elevated volume nature. Data clustering is a data mining technique that exacts the useful data from the data by grouping data into clusters. In big data as the data is complex and of very large volume, individual clustering techniques may not consider all the samples, which may lead to inaccurate results. To overcome this inaccuracy, the proposed method is the combination of dynamic k-means and hierarchical clustering algorithms. This proposed method can be called a hybrid method. Being a hybrid method will overcome a few drawbacks like static k value. In this chapter, the proposed method is compared with existing algorithms by using some clustering metrics.

INTRODUCTION

Big data analytics has become trend in the market and is used to perform analytics on this big data. It is used to extract hidden patterns, unknown correlations and helps organizations in decision making. Big data is the problem and Hadoop is the solution for handling big data available as an open-source framework. Clustering is one of the techniques used to extract insights from big data (Raghupathi & Raghupathi 2014). Traditional clustering techniques may not work for efficient clustering in big data.

DOI: 10.4018/978-1-7998-9640-1.ch015

Consequently, there remains need towards plan an competent & extremely scalable clustering algorithm. This has motivated towards propose a novel algorithm called hybrid clustering algorithm for big data in Hadoop ecosystem (Katal et al., 2013). In Big data analysis characteristics individual clustering techniques like kmeans mean and hierarchical may not consider all the samples which leads to inaccurate results. K-means and hierarchical gathering techniques meet halfway because of the limitations of individual clustering algorithms. Few drawbacks of traditional clustering algorithms are k-means clustering in this algorithm it remains hard towards predict the k value, wrong prediction of k value many data points may not fit into any of the clusters; several merge split decisions and iteration in hierarchical clustering, etc. (Aggarwal & Zhai 2012).

Grouping is important device for information mining & information revelation. The aim of bunching is to discover considerable gatherings of substances moreover to divide groups framed for a dataset. Customary K-implies grouping functions admirably when functional to little datasets (Pandove & Goel 2015). Enormous datasets should be grouped through the end objective that each and all other substance or information point in the bunch is like several elements in a similar group. Grouping issues can be applied to a few bunching disciplines. The capacity towards consequently bunch comparative things empowers one to find covered up likenesses & key ideas while joining a lot of information into a couple of gatherings. This empowers clients towards fathom a lot of information. Groups can be delegated homogeneous & heterogeneous bunches. In homogeneous groups, all hubs contain comparable possessions (Firouzi et al., 2010). Heterogeneous bunches remain exploited in private server farms in which hubs have a variety of attributes moreover in which it could be hard to be familiar with hubs Embrocates (Demchenko et al., 2013).

Clustering techniques require the use of more exact meanings of perception and group likenesses. When gathering depends on ascribes, it is normal to utilize recognizable ideas of distance. An issue with this strategy is related with the estimation of distances between groups including at least two perceptions. (Fernández et al., 2014) In contrast to existing regular measurable techniques, most grouping calculations doesn't depend on factual circulations of information and in this manner can be useful to apply when minimal earlier information exists on a specific issue (Ghazal et al., 2013) portrayed how the quantity of emphases can be diminished by parceling a dataset into covering subsets and by just emphasizing information objects inside covering zones (Battré et al., 2010)

The remainder of this works remains organized as follows. The 'History' section contains relevant surveys on the subject of Big data clustering. We provide a background on Apache Spark in 'Research Paper' The section under 'Study Design' describes the survey's research methods. The section 'Survey Methods' goes through the various Spark clustering algorithms. We provide our analysis on clustering large data with Spark and upcoming projects in 'Discussion and Future Directions.' Lastly, in 'Findings,' bring the paper to be close.

Limitations of Existing Methods

The existing methods like big-data related clustering models with honeybee, genetic and PSO techniques cannot provide accurate bigdata storage. The limitations like static k, dynamic k and hadoop storage issue are cannot solve exactly. The silhouette score, Calinski-Harabasz Index, & Davies - Bouldin Index cannot be improved with this method (Jiang et al., 2010).