

# IMPLEMENTATION OF HIERARCHICAL CLUSTERING AND OUTLIER ANALYSIS IN HADOOP ENVIRONMENT

T Karthikeyan<sup>1</sup>, S Gopinath<sup>2</sup>, M Naveen<sup>3</sup>

<sup>1</sup>Assistant Professor, <sup>2,3</sup>UG Scholars

Department of Computer Science and Engineering

Knowledge Institute of Technology, Salem

Email id: tkcse@kiot.ac.in

## ABSTRACT

Big data is a platform which comprises any considerable amount of structured, semi-structured and unstructured data. Based on the user needs, we need to mine the information as per required format. The main advantage of Big data analytics is predicting the future through the collection of database. The problem behind existing system is how to group similar kind of datasets and to remove unwanted datasets. In order to predict the data, we need to cluster the frequent datasets. To overcome this issue, we proposed the combination of Hierarchical clustering and Outlier analysis algorithm. Hierarchical clustering uses an agglomerative algorithm to make a structured data from unstructured data. In this way, outlier analysis is used to remove the unwanted data. As a result, it enforces easy way to analyze the data quickly and segregate the data.

**Keywords:** Big data, Hierarchical Clustering, Outlier Analysis.

## INTRODUCTION

Big data is a huge amount of data, which cannot store and maintained in a local drive. Therefore it is not efficient and hence need too much computer time to access the data. It is better option to store the huge amount of data [1] in cloud which provide the required amount of memory space. When multiple sources are obtained from the environment, it segregate into respective module [10] and these modules are called as data base. While n numbers of database are collected into a single unit, which is data warehouse. Data mining takes input as a

data warehouse [1] and we get a data for predicting the future

## HIERARCHICAL CLUSTERING

In Hierarchical clustering, an Agglomerative (Bottom up) Algorithm is used to creating a structured data from an unstructured data. Therefore, the organized data is converted into an eloquent cluster. The

output cluster [3] which consist of package of objects which are cognate to each other. The result of a clustering is a tree like structure, which denotes the merging process and buffer cluster is called Dendrogram. In agglomerative bottom up approach method, an individual objects is made as cluster and then continuously join to any one of the most cognate until it find an single cluster [3]. The similarity between the documents can be easily measured by tanimoto, cosine, correlation coefficient & euclidean distance. There are three criterion function and groups are Internal, External, Graph-based, Hybrid. Therefore the traditional selection schema for clustering [8] can be defined as follow; hence both single-link and UPGMA are worst and poor clustering for setting up the data.

1. Single-link,
2. Complete link,
3. UPGMA.

## OUTLIER ANALYSIS

In data analysis large number of data are recorded or analyzed [4]. There exist some data which may be different from the model of the data behavior such different or inconsistent data are called outlier [9] . Although outlier is considered as noise or unwanted things it may contain some valuable information. Outlier considered as important step in a data mining to prevent fraud detection.

These are the ways to detect outlier.

1. Statistical based outlier detection
2. Deviation based outlier detection
3. Distance based outlier detection
4. Statistical based outlier detection

The most efficient method to detect outlier detection is classified into two types

1. Distribution based outlier detection
2. Depth based outlier detection

Outlier detection is the preliminary process in the data mining. Hence this can be calculated using the distance measure, clustering and spatial method [6]. When the data obtained is purely clustered with similarity of datum, then it need to perform the statistical based outlier analyse [7] to remove the unwanted data.

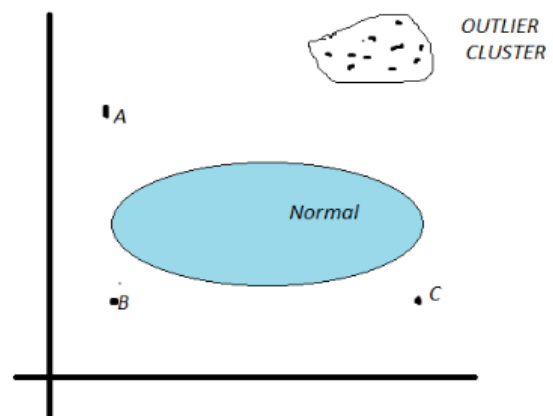


Fig1: Outlier Cluster

## IMPLEMENTATION OF HIERARCHICAL CLUSTERING &

## OUTLIER ANALYSIS IN HADOOP ECOSYSTEM

When the data is collected from an multiple sources of an environment, it is necessary to group (i.e.) clustering using the hierarchical algorithm [8] through the bottom up approach .where an individual object is made as cluster and then continuously joined with most similar objects until it end.

However similarity between the documents can be measured easily using cosine.

$$\cos(d_i, d_j) = \frac{d_i^t d_j}{\|d_i\| \|d_j\|}$$

If the source data is found then the Hybrid criterion functions tries to found the various possibilities and measures of similarity over the documents in an each cluster [9] and also minimize the possibilities of similarities between each cluster.

There are two hybrid criterion functions. First one is obtained by with

$$\text{maximize } H_1 = \frac{I_1}{\varepsilon_1} = \frac{\sum_{r=1}^k \|D_r\|^2 / n_r}{\sum_{r=1}^k n_r D_r^t D / \|D_r\|}$$

Therefore second law is obtained by, with

$$\text{maximize } H_2 = \frac{I_2}{\varepsilon_1} = \frac{\sum_{r=1}^k \|D_r\|}{\sum_{r=1}^k n_r D_r^t D / \|D_r\|}$$

Both the Single-link and UPGMA scheme are very poor in case of finding the distance between the two clusters. Single-link which suffers on chaining effect during clustering, in addition to that it produces the clusters which are similar to noisy pattern.

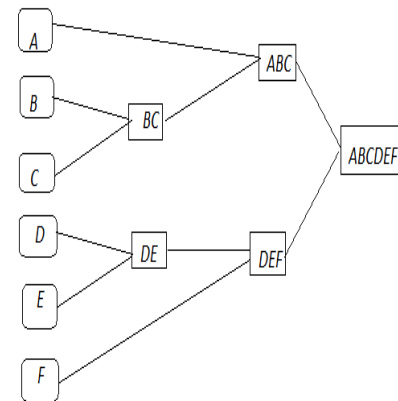


Fig2: Multi-source data ,Clustering of similar data,  
Grouped data

So it is better option to use complete-link scheme. Because, this algorithm produces tightly bound or compact cluster [2] which is difficult to extract the concentric clusters. Which is most accurate and economic.

## CONCLUSION

By using hierarchical clustering in hadoop, it can be able to cluster the similar type of data. The hierarchical clustering first divides the data into partition and group the similar type of data. The outlier analysis is used to discarded the noisy data and produce the cluster. It can be effectively utilized in hadoop environment. The time efficiency can be reduced to process the large datasets.

## REFERENCES

- [1] Jiawei Han, MichelineKamber,"Data Mining Concepts and Techniques", Second Edition, Morgan Kaufman Publishers, 2006 Elsevier.
- [2] Konstantin Shvachko, HairongKuang, "The Hadoop Distributed File System", Published by IEEE, 2010.
- [3] Y. Luo, Z. Guo, Y. Sun, B. Plale, J. Qiu, and W. W. Li, "A hierarchical framework for cross-domain mapreduce execution,"in Proceedings of the second international workshop on Emerging computational methods for the life sciences, ser. ECMLS '11. New York, NY, USA: ACM, 2011, pp. 15–22.
- [4] Anscombe, F. J. and Guttman, I. 1960. "Rejection of outliers. Technometrics "2, 2, 123 - 147. Arning.
- [5] A.,Agrawal, R., and Raghavan, P. 1996. "A linear method for deviation detection in large databases". In Proceedings of 2nd International Conference of Knowledge Discovery and Data Mining. 164 – 169.
- [6] Torr, P. and Murray, D. 1993. "Outlier detection and motion segmentation". In Proceedings of SPIE, Sensor Fusion VI, Paul S. Schenker; Ed. Vol. 2059. 432 - 443.
- [7] P. Olmo Vaz de Melo, L. Akoglu, C. Faloutsos, and A. "Loureiro.Surprising Patterns for the Call Duration Distribution of Mobile Phone Users", ECML/PKDD Conference, 2010.
- [8] Zhang Zhen Ya, Cheng Hong Mei, Wang Jin, Wang Xu Fa, "An Approach on the Data Structure for the Matrix Storing Based on the Implementation of Agglomerative Hierarchical Clustering Algorithm", Computer Science, 2006(1), pp.14-17.
- [9] Jiangtao Yin —Accelerating Expectation-Maximization Algorithms with Frequent Updates□ 2012 IEEE International Conference on Cluster Computing.

[10]Yahoo! Launches World's Largest  
Hadoop Production Application,  
<http://tinyurl.com/2hgzy7>.