ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Special Issue 14, April 2016.

IMPLEMENTATION OF PATTERN DETECTION AND OUTLIER ANALYSIS IN HADOOP ECOSYSTEM

T Karthikeyan[1], R Rekhaa[2]

[1]Research Scholar, GITAM University, Visakhapatnam, tkcse@kiot.ac.in
[2]Department of Computer Science and Engineering, Knowledge Institute of Technology, Salem, rekhaaramesh1997@gmail.com

ABSTRACT

Big Data is a broad term of data sets of large and complex (ie) a huge amount of data are called Big Data. Challenges in big data include analysis, capture, data curation, search, sharing, storage, transfer and visualization. In the proposed work, we are going to make use of two algorithms such as Pattern Detection and Outlier Analysis. Pattern Detection: Generally pattern Detection is grouping similar number of data to form cluster. By combining Outlier Analysis with Pattern Detection, we can reduce the missing values. It may be due to variability in the measurement or it may indicate Experimental error. The remaining data which contains noise or missing values can be excluded from the data set by using Outlier Analysis. By implementing the above algorithms in the platform of big data, the characteristics of massive amount of data can be detected, we can control the time management at the same time frequent data and missing values can be easily found and can perform task at a given time. Finally a noise made in the analysis of outlier can be reduced by implementing these two algorithms in the platform of Big Data.

Keywords: Big Data, Pattern Detection and Outlier Analysis.

INTRODUCTION

Data Mining which is a process of extracting required data from a huge amount of data (Data Warehouse) [1]. By implementing Pattern Detection in the platform of Data Mining, we can group the similar data to form cluster. For example: Assume that faculty or students bio-data from various departments are stored in the main office in the college. Now similar data (ie) same department students and faculty bio-data have to be clustered [2]. By forming cluster and by extracting a data from a huge amount of data, missing values can occur. To overcome this, Outlier Analysis can be implementing. This Outlier Analysis helps us to reduce missing values [6] and a massive amount of data can be detected. If the user allocates a certain time to perform one task, Outlier Analysis algorithm [8] helps in time management at the same time frequent amount of data. Outlier Analysis produces noise, while performing task. By combining Outlier Analysis and Pattern Detection [1] , we can achieve grouping of similar data sets and removal of unwanted data sets. As a result,

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Special Issue 14, April 2016.

these two algorithms help us to solve several problems in the concept of Data Mining.

PATTERN DETECTION

Generally Pattern Detection means grouping similar number of data to form cluster.A Number of methods for extracting meaning from data sets through a combination of operations like research methods, a graph theory,data analysis,clustering and advanced mathematics.By implementing Pattern Detection in the Hadoop[2] environment massive amount of data can be detected,without making use of graph representation for a given data set we can approach to data analysis,without allowing ranking algorithm[10] we can quantify the pattern in the graph and the way of approach becomes effective for images,audio and video which were commonly applied in machine learning techniques [7] in the analyzes of numerical and text data.On the whole, we can segregate massive amount of data and we can analyze numerical and text data by implementing Pattern Detection [3] in the platform of Hadoop.

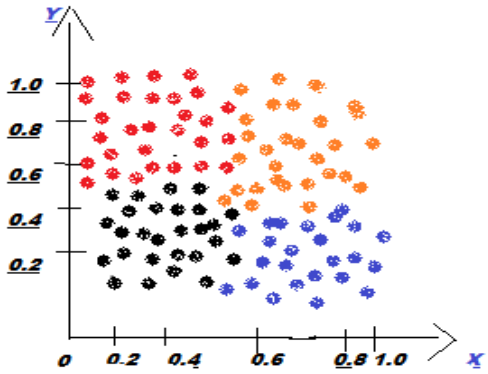

Fig 1: Clustering of Data in Hadoop Environment

OUTLIER ANALYSIS

It is an observation point that is distant from other observation. Sometimes it may exclude from the data set. It may be due to variability in the measurement or it may indicate experimental error.By implementing Outlier Analysis in the platform of Big Data [9], we can control the time management at the same time frequent amount of data and the missing values can be easily found and can perform task at the given time. Finally a noise made in the analysis of outlier [7] can be reduced by implementing this algorithm in the platform of Big Data.

PATTERN DETECTION AND OUTLIER ANALYSIS IN HADOOP

In Online shopping, the websites like flipkart, Amazon, my price etc. The products in this websites are categorized into different ranges like high price, relevance, low cost, new trends. On this above separation of products, a similar amount of

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Special Issue 14, April 2016.

product are grouped together to form cluster and massive amount of products can be detected [5] by the implementation of Pattern Detection. But on this separation of products based on category, missing values can occur and a noise can be produced. To overcome this an algorithm called outlier Analysis [4] is implemented in this platform. By using this, we can control the time management, missing values can be easily found and a noise produced can be reduced. In addition to that frequent amount of products [6] can be found and it can perform multiple tasks at a given time.

ADVANTAGES

- Using proposed system, frequent data sets can be identified.
- HDFS-By making use of map reduce concept we can achieve better result.

CONCLUSION AND FUTURE ENHANCEMENT

To conclude,using Pattern Detection and Outlier Analysis,a missing values can be detected and a noise produced can be reduced.In Future by using K-means algorithm and Apriori algorithm a frequent amount of data can be grouped in the form of cluster in the efficient manner.

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] Mine Olson, "Hadoop scalable, flexible data storage and analysis", Vol1 no 3.

[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] M. Otey, S. Parthasarathy, and A. Ghoting." Fast Distributed OutlierDetection in Mixed Attribute Data Sets, Data Mining and Knowledge Discovery", 12(2–3), pp. 203–228, 2006.

[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/2hgzv7.