An Inception towards better Big Data Clustering Technique

P S.Md.Mujeeb, R.Praveen Sam, K.Madhavi

Abstract--- The speedy emanating technology during past few year in the area of information technology is "Big Data". Clustering is one of the crucial task in broad range of domains handling enormous data. This survey presents the various clustering approaches adopted for the effective big data clustering. Thus, this review article provides the review of 15 research papers suggesting various methods adopted for the effective big data clustering, like K-means clustering, Variant of K-means clustering, Fuzzy C-means clustering, Possibilistic C-means clustering, Collaborative filtering and Optimization based clustering. Moreover, an elaborative analysis is done by concerning the implementation tools used, datasets utilized and the adopted framework for clustering of big data. Subsequently an effective scheme must be developed to surpass present techniques for exceptional management of big data. Eventually the research issues and gaps of various big data clustering techniques are presented for benefiting the researchers for inception towards better big data clustering.

Keywords: Big data, MapReduce, clustering, K-mean, C-mean.

1. INTRODUCTION

Over the past few years the "Big data" has grown as one of the alluring industry of the information technology sector. The Big data is a term generally utilized for emphasizing the challenges and advantages encounter during the collection and handling of enormous amount of data [1]. The actual definition of Big data, is the amount of data which outpace the processing capacity of a particular system in terms of consumption of time and memory utilization. The big data has attracted the interest of vast range of fields like retail, financial businesses, e-commerce, medicine and various industries who are handling enormous amount of raw data. Anyhow the process of analysis and acquiring of knowledge from big data is becoming problematic in all most all basic and advanced data mining tools [2]. Clustering is a important method utilized in the area of knowledge discovery and data engineering. The main objective of clustering is to gather the given data or objects into a distinct group of objects in accord with their special metrics for grouping the objects into the homogeneous group. There are complications in applying clustering methods to big data because of the fresh challenges that are inflated with big data [3]. This paper's primary intention is to provide survey of various big data clustering methodologies for the effective management of big data. The survey is made by considering the implementation tools used, datasets utilized and framework adopted for clustering of big data and the additional survey was done to exploit the research gaps and issues. Hence, an inception for better and efficient big data clustering technique.

This article is arranged as: Section 1 providing brief introduction about this article, Section 2 gives the literature review of present big data clustering schemes, Section 3 briefs about the research gaps and issues identified, in Section 4 elaborates the analysis of various tools and framework used and the conclusion of this article is made in Section 5.

2. LITERATURE SURVEY ON VARIOUS BIG DATA CLUSTERING SCHEMES

This section discusses the review of the various research papers employed for big data clustering methodologies for the compelling big data management. The classification of distinct big data clustering methods are shown in Figure 1. They are K-means clustering, Variant of K-means clustering, Fuzzy C-means clustering, Possibilistic C-means clustering, Collaborative filtering and Optimization based clustering.

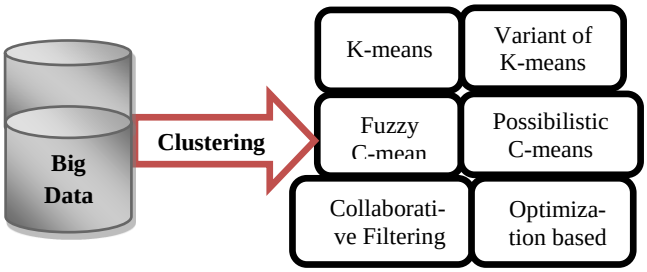

Figure 1. Categorization of distinct big data clustering schemes.

a) K-means clustering

The different research papers utilizing the K-means clustering for the big data are elaborately discussed below,

Sreedhar C. *et al.* [4] proposed a K-Means Hadoop MapReduce (KM-HMR) for the effective big data clustering. In this two approaches were presented for MapReduce (MR) framework based clustering. The KM-HMR was the first method, which concentrated on the application of MR on standard K-means. The second method was to enhance the clusters quality by minimizing the inter-cluster distances and maximizing intra-cluster distances. The proposed KM-HMR approaches results have outperformed the efficiency of other clustering methods with respect to execution time.

Revised Manuscript Received on December 22, 2018.

P S.Md.Mujeeb, Research Scholar, JNTUA, Ananthapuramu-515002, Andhra Pradesh, India (mujeeb.smd@gmail.com)

Dr. R.Praveen Sam, Professor, Department of CSE, GPREC, Kurnool-518007, Andhra Pradesh, India (praveen_sam75@yahoo.com)

Dr. K.Madhavi, Associate Professor, Department of CSE, JNTUACEA, Ananthapuramu-515002, Andhra Pradesh, India (kasamadhavi@yahoo.com)

Nadeem Akthar *et al.* [5] recommended a altered K-means clustering algorithm, which selects the K-optimal data points in the dataset. The main advantage of selecting the data points from the massive datasets is to prevent outlier points from involving in the final evaluation of the cluster. More stable results were attained when the initial centres were sorted for appropriate datasets.

Ankita S. and Prasanta K. Jana [6] proposed a K-means clustering algorithm implemented in Spark. The proposed K-Means algorithm solved the resolution issues which is present in the common K-Means clustering algorithm by prior automation of the input clusters. It resulted in better performance of the Spark framework based K-means clustering algorithm even with the increased size of the data and even machine count.

b) Variant of K-means clustering

The different research papers utilizing the Variant of K-means clustering for the big data are elaborately discussed below,

Mohamed Aymen Ben HajKacem *et al.* [7] proposed the Accelerated MapReduce-based K-Prototypes (AMRKP) clustering methodology for handling big data. In this method the reading & writing of data were done only once because of this the number of Input and Output (I/O) operations were reduced drastically. Furthermost, the proposed scheme is dependent on pruning strategy to accelerate the process of clustering by minimization the redundant distance between the centre and data points of the cluster. The developed AMRKP surpass the other clustering schemes with respect to efficiency and scalability.

M. Omair Shafiq and Eric Torunski [8] proposed a parallel K-Medoids clustering algorithm based on MR framework for carrying out the effective clustering of huge database. The devised clustering method was efficient, simple and capable of handling the datasets with fluctuating varying characteristics, like volume, velocity and variety. The simulation result displayed the capability and feasibility of proposed clustering method in handling large scale datasets.

Mohamed Aymen Ben Haj Kacem *et al.* [9] proposed an MR framework using K-Prototypes (MR-KP) clustering scheme for the effective data clustering using parallelization. It resulted in being a famous and effective clustering method for mixed massive datasets. The simulation was performed on millions of samples and the outcome was accurate and scalable even when the size of data is increased.

c) Fuzzy C-means (FCM) clustering

The different research papers utilizing the FCM clustering for the big data are elaborately discussed below,

Simone A Ludwig [10] investigated the scalability and parallelization of FCM clustering algorithm. The FCM clustering algorithm was parallelized using MR framework by outlining the procedure of map and reduce function. The validation analysis of the MR-FCM clustering algorithm was made to show the effectiveness of the proposed algorithm with respect to the aspect of purity.

Minyar Sassi Hidri *et al.* [11] proposed an enhanced FCM clustering algorithm using sampling mixed with split and merge strategy for clustering big data . Initial step is to split data into distinct subsets and operate the individual nodes in parallely. Then, the subsets were sampled, which were again split randomly into distinct subsamples. This algorithm performed effectively with the provided resources with optimized time and space complexities.

d) Possibilistic C-means (PCM) clustering

The different research papers utilizing the PCM clustering for the big data are elaborately discussed below,

Qingchen Zhang and Zhikui Chen [12] suggested a weighted kernel PCM algorithm (wkPCM) for clustering the data objects into the suitable groups. The kernel weights were integrated for defining the object's importance in the kernel clustering for minimizing the corruption generated by the noisy data. The proposed distributed wkPCM clustering algorithm was based on the MR framework which can equip compelling computational speed for real time data sets.

Qingchen Zhang *et al.* [13] proposed a Privacy-Preserving High-Order PCM (PPHOPCM) clustering algorithm for performing the clustering of big data through the optimizing the objective function. The distributed HOPCM method is based on the MR framework with the aim of dealing with big data. Eventually, the PPHOPCM was used for protecting the data on cloud by applying the Brakerski-Gentry-Vaikuntanathan (BGV) encryption scheme on HOPCM. In the PPHOPCM, membership matrix and cluster centres were updated using polynomial functions. The devised PPHOPCM algorithm effectively clustered the huge dataset and secure the private cloud data.

e) Collaborative Filtering (CF) based clustering

The different research papers utilizing CF clustering for the big data are elaborately discussed below,

Rong Hu *et al.* [14] designed a Clustering-based Collaborative Filtering (ClubCF) approach whose intension is to provide similar services recruitment in the same clusters for the recommendation of services collaborative. This approach of clustering is comprised of two phases. In the first phase the data sets are decomposed into small chunks of clusters to make them suitable for next processing. In the next phase CF is applied to the determined clusters. Since the number services count in the cluster was less than the available web services the time complexity of CF was lesser comparatively.

Subramaniyaswamy V. *et al.* [15] proposed the predictive mechanism based CF method for the effective processing of large-scale data parallely. The MR framework is used for carrying out the aggregation, filtering and maintenance of the efficient storage. The CF was used to refining of data. The developed clustering scheme was improved by processing the input data into emoticon and tokens through the application of sentiment analysis. The simulation resulted in significant improvement in the complexity analysis performance.

f) Optimization-based clustering

The different research papers utilizing the optimization-based clustering for the big data are elaborately discussed below,

P. Sachar and V. Khullar [16] presented a Genetic Algorithm (GA) based clustering algorithm utilizing Hadoop MR framework. GA with the help of Hadoop resulted in better execution time and space complexities that in-turn reduced the user cost for clustering. The main asset of the GA was its effectiveness and iterative nature; hence it was opted on above java based GA.

J. Karimov and M. Ozbayoglu [17] developed a Hybrid Evolutionary Clustering with Empty Clustering Solution ($H(EC)^2S$) by integrating the Fireworks and Cuckoo Search (CS) algorithm with some heuristics related to centroid-calculation. Initially, to eliminate empty clustering issues some representative points were selected. Then these points were used by the hybrid algorithm for the selection of centroid. The main advantage of the proposed method is particularly when the number, amount and dimensionality of cluster parameters likely to increase.

Yan Yang *et al.* [18] designed a Semi-supervised Multi-Ant Colonies clustering algorithm implementing on the Hadoop MR framework. This clustering methodology was developed to deal with massive data sets or big data. This method incorporated the pair-wise constraints in each and every ant colony clustering process and estimates the new similarity matrix. The proposed clustering algorithm has improved the computational efficiency of the big data clustering through the help of the MR framework.

3. RESEARCH GAPS IDENTIFIED

This section deals with the various research gaps and issues in different big data clustering methodologies.

The proposed K-Means Hadoop MapReduce (KM-HMR) clustering method was not able to provide satisfactory job scheduling for huge datasets in turn degrading the performance of map and reduce for large datasets [4]. In the proposed K-means clustering algorithm the distance calculation function is very complex and time exhausting process [5]. The devised K-means clustering algorithm implemented in Apache Spark framework didn't consider few important factors of big data such as veracity and velocity, and moreover this algorithm couldn't effectively cluster the real time streaming big data [6]. In the employed variant of K-means clustering method the modelled Accelerated MapReduce-based K-Prototypes (AMRKP) was unable to reduce the iteration count and improve the scalability [7]. The proposed Parallel K-Medoids Algorithm for big data clustering was not applicable to clusters of large scale data intensive applications [8]. In the devised MapReduce-based K-Prototypes (MR-KP) clustering method parallelizing in the initialization step to create set of cluster center is missing and this method was not applicable to real time applications such as fraud detections [9]. MapReduce fuzzy c-mean (MR-FCM) clustering algorithm was not efficient enough for larger data sets comprising of Gigabytes of data [10]. The devised Fuzzy c-mean clustering algorithm has to exploit how to extract frequent item sets effectively as it is essential step for data analysis [11]. In the proposed weighted kernel Possibilistic c-Means (wkPCM) algorithm for clustering big data the analysis of multi-dimensional space and deep features are missing [12]. High-order Possibilistic c-Mean (HOPCM) scheme's efficiency can be improved with the help of cloud servers for high scalable big data clustering [13]. The Clustering-based Collaborative Filtering (ClubCF) approach was not able to resolve the scarcity issues which can be enhanced by semantic analysis for better coverage [14]. The used apache mahout Collaborative filtering technique for recommendation generation process was not efficient enough and can be still optimized further for big data clusters [15]. Because of the optimization nature the approach that employed genetic algorithm for big data clustering can still improve the efficiency of Big data Analytics [16]. The proposed hybrid evolutionary clustering model's has a runtime disadvantage when compared with other algorithms [17]. In the devised parallel ant colony clustering method the overhead of iteration is overwhelming which is effecting the performance of algorithm [18].

4. ANALYSIS & RESULTS

The analysis the distinct research work for the clustering of big data with respect to implementation tool, dataset utilized and framework adopted are discussed in this section.

4.1 Analysis based on Implementation tool

This subsection tells about the analysis carried out considering the implementation tool employed in the above mentioned research papers. Table 1 displays the various implementation tools employed for the effective clustering of the big data. The commonly used tools for big data clustering are Java, Cloudera, VC++ and R programming. From the Table 1, it is clear that Java is the most frequently employed implementation tool for the effective clustering of big data.

Table 1. Analysis with respect to the implementation tool used in clustering approaches.

Implementation Tools	Research papers
Java	[4] [5] [8] [9] [10] [14] [15] [16]
Cloudera	[12] [13] [17]
VC++	[18]
R Programming	[11]

4.2 Analysis based on Datasets utilized

This subsection tells about the analysis carried out with respect to the datasets utilized by the various above mentioned research works.

The familiar datasets utilized in the different research works are KDD Cup'99 dataset, Covertype dataset, Pokerhand dataset, Iris dataset, Susy dataset, Wine dataset and Synthetic dataset. From the Figure 2, it is reflected that the most frequently utilized datasets are KDD Cup'99 and Iris datasets.

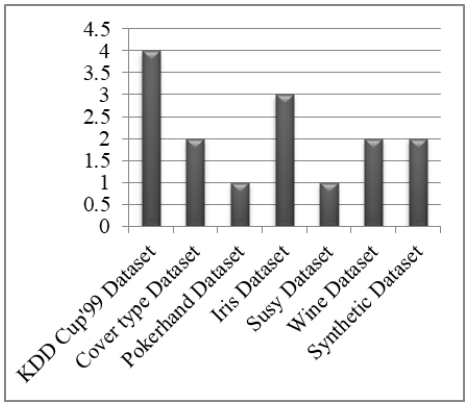

Figure 2. Analysis based on dataset utilized in clustering approaches.

4.3 Analysis based on the framework adopted

This subsection analysis is carried out by with respect to the various frameworks adopted for big data clustering. Table 2 shows different frameworks utilized for big data clustering. From the Table 2 we can visualize that MapReduce framework is the most commonly used framework for treating big data clustering.

Table 2. Analysis based on framework adopted in clustering schemes

Frameworks	Research papers
MapReduce Framework	[7] [8] [9] [10] [11] [12] [15]
Apache Spark Framework	[6]
Hadoop Framework	[4] [14] [17]
Hadoop MapReduce Framework	[5] [16] [18]

5. CONCLUSION

A survey on the different clustering schemes employed for the effective clustering of the big data is explained in this paper. The main intention of this article is to study, learn and categorize the distinct clustering techniques utilized for the big data by analysis of 15 research papers from IEEE, Elsevier, Springer, Google Scholar and various International journals. The analysis were made concerning the adopted implementation tools, utilized datasets and employed frameworks. This survey also suggests the major future scope for the inception of effective big data clustering by considering the issues and research gaps briefed in Section3. In conformity with the analysis and discussion, it can be concluded that Java is the frequently used implementation tool and MapReduce is the vastly adopted framework for effective clustering of big data.

REFERENCES

1 Wullianallur Raghupathi, and Viju Raghupathi, "Big data analytics in healthcare: promise and potential", Health information science and systems, vol. 2, no. 1, pp. 3, 2014.

2 Bhagyashri S. Gandhi, and Leena A. Deshpande, "The survey on approaches to efficient clustering and classification analysis of big data", International Journal of Engineering Trends and Technology (IJETT), vol. 36, no. 1, pp. 33-39, 2016.

3 Ali Seyed Shirkhorshidi, Saeed Aghabozorgi, Teh Ying Wah and Tutut Herawan, "Big Data Clustering: A Review", Computational science and its applications - ICCSA 2014: 14th international conference Guimarães, Portugal, june 30 - july 3, 2014 proceedings.

4 Chowdam Sreedhar, Nagulapally Kasiviswanath, and Pakanti Chenna Reddy, "Clustering large datasets using K-means modified inter and intra clustering (KM-I2C) in Hadoop", Journal of Big Data, vol. 4, no. 1, pp. 27, 2017.

5 Nadeem Akthar, Mohd Vasim Ahamad, and Shahbaz Khan, "Clustering on Big Data Using Hadoop MapReduce", in proceedings of 2015 IEEE International Conference on Computational Intelligence and Communication Networks (CICN), pp. 789-795, 2015.

6 Ankita Sinha, and Prasanta K. Jana, "A novel K-means based clustering algorithm for big data", in proceedings of 2016 IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1875-1879, 2016.

7 Mohamed Aymen Ben HajKacem, Chiheb-Eddine Ben N'cir, and Nadia Essoussi, "One-pass MapReduce-based clustering method for mixed large scale data", Journal of Intelligent Information Systems, pp.1-18, 2017.

8 M. Omair Shafiq, and Eric Torunski, "A Parallel K-Medoids Algorithm for Clustering based on MapReduce", in proceedings of 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 502-507, 2016.

9 Mohamed Aymen Ben Haj Kacem, Chiheb-Eddine Ben N'cir, and Nadia Essoussi, "MapReduce-based k-prototypes clustering method for big data", in proceedings of 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 1-7, 2015.

10 Simone A Ludwig, "MapReduce-based fuzzy c-means clustering algorithm: implementation and scalability", International Journal of Machine Learning and Cybernetics, vol. 6, no. 6, pp. 923-934, 2015.

11 Minyar Sassi Hidri, Mohamed Ali Zoghlami, and Rahma Ben Ayed, "Speeding up the large-scale consensus fuzzy clustering for handling Big Data", Fuzzy Sets and Systems, 2017.

12 Qingchen Zhang, and Zhikui Chen, "A weighted kernel possibilistic c-means algorithm based on cloud computing for clustering big data", International Journal of Communication Systems, vol. 27, no. 9, pp. 1378-1391, 2014.

13 Qingchen Zhang, Laurence T. Yang, Zhikui Chen, and Peng Li, "PPHOPCM: Privacy-preserving High-order Possibilistic c-Means Algorithm for Big Data Clustering with Cloud Computing", IEEE Transactions on Big Data, vol. pp, no. 99, pp. 1-11, 2017.

14 Rong Hu, Wanchun Dou, and Jianxun Liu, "ClubCF: A clustering-based collaborative filtering approach for big data application", IEEE transactions on emerging topics in computing, vol. 2, no. 3, pp. 302-313, 2014.

15 V. Subramaniyaswamy, V. Vijayakumar, R. Logesh, and V. Indragandhi, "Unstructured data analysis on big data using MapReduce", Procedia Computer Science, vol. 50, pp. 456-465, 2015.

16 P. Sachar and V. Khullar, "Social media generated big data clustering using genetic algorithm", in proceedings of 2017 IEEE International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, pp. 1-6, 2017.

17 Jeyhun Karimov, and Murat Ozbayoglu, "High quality clustering of big data and solving empty-clustering problem with an evolutionary hybrid algorithm", in proceedings of 2015 IEEE International Conference on Big Data (Big Data), pp. 1473-1478, 2015.

18 Yan Yang, Fei Teng, Tianrui Li, Hao Wang, Hongjun Wang, and Qi Zhang, "Parallel Semi-supervised Multi-Ant Colonies Clustering Ensemble Based on MapReduce Methodology", IEEE Transactions on Cloud Computing, vol.6, no.1, pp. 1-12, 2015.