International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-6, April 2020

Analysis towards Enhanced Big Data Clustering Technique

Jagdish Kushwaha, Shailesh Jaloree, R.S.Thakur

Abstract--- The expedient exuding innovation during recent year in the zone of data innovation is "Huge Data". Grouping is one of the significant assignment in wide scope of areas dealing with gigantic information. This study presents the different bunching approaches received for the viable enormous information grouping. Therefore, this survey article gives the audit of various research papers proposing different strategies embraced for the successful huge information grouping, similar to K-implies bunching, Variant of K-implies bunching, Fuzzy C-implies grouping, Possibilistic C-implies bunching, Collaborative separating and Optimization based bunching. In addition, an elaborative examination is finished by concerning the usage instruments utilized, datasets used and the received system for bunching of huge information. In this manner a successful plan must be created to outperform present systems for remarkable administration of enormous information. In the long run the examination issues and holes of different huge information bunching strategies are introduced for profiting the analysts for initiation towards better large information grouping.

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

I. INTRODUCTION

In the course of recent years the "Large information" has developed as one of the charming business of the data innovation part. The Large information is a term commonly used for stressing the difficulties and favorable circumstances experience during the assortment and treatment of huge measure of information [1]. The genuine meaning of Enormous information, is the measure of information which outpace the preparing limit of a specific framework as far as utilization of time and memory use. The huge information has pulled in light of a legitimate concern for immense scope of fields like retail, money related organizations, online business, medication and different ventures who are taking care of tremendous measure of crude information. At any rate the procedure of examination and securing of information from enormous information is getting risky in all most all essential and propelled information mining apparatuses [2]. Bunching is a significant strategy used in the subject matter revelation and information building.

Revised Manuscript Received on April 30, 2020.
***** Correspondence Author
Jagdish Kushwaha* is currently pursuing Ph.D. degree program in computer applications in BU Bhopal, ,+919630831212. E-mail: mca.jagdish@gmail.com
Dr.(Prof) Shailesh Jaloree ,Professor department of computer science & applied mathematics SATI VIDISHA(MP),9424418356.shailesh_jaloree@rediffmail.com.
Dr (Prof) R.S.THAKUR ,Professor and Department of mathematics, Bioinformatics & Computer Application MANIT Bhopal(MP) 9826241996.ramthakur2000@gmail.com.

The fundamental goal of bunching is to assemble the given information or articles into an unmistakable gathering of items as per their exceptional measurements for gathering the items into the homogeneous gathering. There are inconveniences in applying grouping techniques to enormous information in view of the new difficulties that are swelled with large information [3]. This present paper's essential aim is to give study of different large information grouping systems for the compelling administration of huge information.

The review is made by considering the execution devices utilized, datasets used and structure received for grouping of huge information and the extra overview was done to misuse the examination holes and issues. Henceforth, a beginning for better and proficient enormous information bunching strategy.

II. 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.

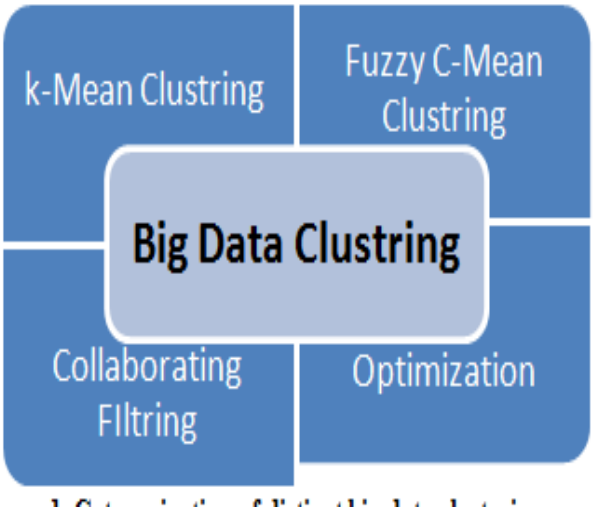

Figure 1. Categorization of distinct big data clustering schemes.

a) K-means clustering

The distinctive research papers using the K-implies bunching for the huge information are intricately talked about underneath, Sreedhar C. et al. [4] proposed a K-Means Hadoop MapReduce (KM-HMR) for the successful large information bunching.

Right now were introduced for MapReduce (MR) system based grouping. The KM-HMR was the primary technique, which focused on the use of MR on standard K-implies. The subsequent strategy was to improve the bunches quality by limiting the between group removes and amplifying intra-group separations. The proposed KM-HMR approaches results have beated the effectiveness of other grouping strategies as for execution time. Nadeem Akthar et al. [5] prescribed a changed K-implies grouping calculation, which chooses the K-ideal information focuses in the dataset. The principle favorable position of choosing the information focuses from the gigantic datasets is to forestall anomaly focuses from including in the last assessment of the bunch. Increasingly steady outcomes were accomplished when the underlying focuses were arranged for proper datasets. Ankita S. furthermore, Prasanta K. Jana [6] proposed a K-implies bunching calculation executed in Sparkle. The proposed K-Means calculation explained the goals issues which is available in the basic K-Means bunching calculation by earlier robotization of the information groups. It brought about better execution of the Flash structure based K-implies grouping calculation even with the expanded size of the information and even machine check. the size of information is expanded.

b) Fuzzy C-means (FCM) clustering

The distinctive research papers using the FCM bunching for the enormous information are intricately talked about underneath, Simone A Ludwig [10] explored the versatility and parallelization of FCM grouping calculation. The FCM bunching calculation was parallelized utilizing MR structure by sketching out the methodology of guide and lessen work. The approval examination of the MR-FCM grouping calculation was made to show the viability of the proposed calculation concerning the part of immaculateness.

Minyar Sassi Hidri et al. [11] proposed an improved FCM grouping calculation utilizing inspecting blended in with split and information into unmistakable subsets and work the individual hubs in parallely. At that point, the subsets were examined, which were again part haphazardly into unmistakable subsamples. This calculation performed viably with the furnished assets with streamlined reality complexities.

c) Collaborative Filtering (CF) based clustering

The diverse research papers using CF grouping for the large information are extravagantly examined beneath, Rong Hu et al. [14] planned a Bunching based Community oriented Sifting (ClubCF) approach whose intension is to offer comparable types of assistance enlistment in similar groups for the suggestion of administrations collective. This methodology of grouping is involved two stages. In the main stage the informational collections are deteriorated into little lumps of groups to make them appropriate for next handling. In the following stage CF is applied to the decided groups. Since the number administrations include in the group was not exactly the accessible web benefits the time multifaceted nature of CF was lesser relatively.

Subramaniyaswamy V. et al. [15] proposed the prescient component based CF technique for the compelling handling of enormous scope information parallely. The MR structure is utilized for completing the collection, sifting and upkeep of the proficient stockpiling. The CF was accustomed to refining of information. The created grouping plan was improved by handling the information into emoji and tokens through the use of conclusion examination. The recreation brought about critical improvement in the multifaceted nature examination execution.

d) Optimization-based clustering

The distinctive research papers using the improvement based bunching for the enormous information are intricately talked about underneath, P. Sachar and V. Khullar [16] introduced a Hereditary Calculation (GA) based grouping calculation using Hadoop MR structure. GA with the assistance of Hadoop brought about better execution existence complexities that thus decreased the client cost for grouping. The primary resource of the GA was its adequacy and iterative nature; thus it was selected on above java based GA.

J. Karimov and M. Ozbayoglu [17] built up a Cross breed Developmental Grouping with Void Bunching Arrangement (H (EC)2 S) by incorporating the Firecrackers and Cuckoo Search (CS) calculation with certain heuristics identified with centroid-count. At first, to take out void grouping issues some agent focuses were chosen. At that point these focuses were utilized by the half and half calculation for the choice of centroid. The primary bit of leeway of the proposed strategy is especially when the number, sum and dimensionality of group parameters prone to increment.

Yan Yang et al. [18] structured a Semi-managed Multi-Subterranean insect Settlements grouping calculation actualizing on the Hadoop MR system. This bunching system was created to manage huge informational indexes or enormous information. This technique consolidated the pair-wise requirements in every single subterranean insect settlement bunching procedure and assessments the new closeness network. The proposed grouping calculation has improved the computational proficiency of the enormous information bunching through the assistance of the MR system.

III. RESEARCH GAPS IDENTIFIED

This segment manages the different research holes and issues in various enormous information bunching strategies. The proposed K-Means Hadoop MapReduce (KM-HMR) grouping strategy was not ready to give acceptable occupation booking to enormous datasets thusly debasing the presentation of guide and decrease for huge datasets [4]. In the proposed K-implies grouping calculation the separation computation work is perplexing and time debilitating procedure [5]. The conceived K-implies bunching calculation actualized in Apache Sparkle structure didn't consider hardly any significant elements of huge information, for example, veracity and speed, and also this calculation couldn't adequately group the continuous spilling huge information [6]. In the utilized variation of K-implies grouping strategy the demonstrated Quickened MapReduce-based K-Models (AMRKP) couldn't decrease the cycle tally and improve the versatility [7]. The proposed Equal K-Medoids Calculation for enormous information bunching was not relevant to groups of huge scope information escalated applications [8]. In the formulated MapReduce-based K-Models (MR-KP) bunching strategy parallelizing in the introduction venture to make set of group focus is missing and this technique was not pertinent to continuous applications, for example, misrepresentation recognitions [9].

MapReduce fluffy c-mean (MR-FCM) bunching calculation was not effective enough for bigger informational indexes containing Gigabytes of information [10]. The formulated Fluffy c-mean bunching calculation needs to misuse how to remove visit thing sets adequately as it is basic advance for information examination [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].

IV. 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.

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.

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.

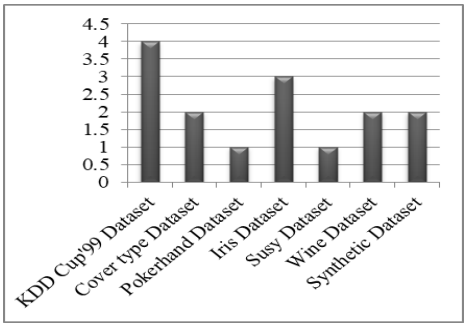

Figure 2. Analysis based on dataset utilized in clustering approaches.

Analysis based on the framework adopted

This subsection analysis is carried out by with respect to the various frameworks adopted for big data clustering. different frameworks utilized for big data clustering.

V. 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 various 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.

AUTHORS PROFILE

Jagdish Kushwaha is currently pursuing Ph.D. degree program in computer applications in BU Bhopal, ,+919630831212. E-mail: mca.jagdish@gmail.com

.

Dr.(Prof) Shailesh Jaloree ,Professor & HOD department of computer science & applied mathematics SATI Vidisha(MP),9424418356.shailesh_jaloree@rediffmail.com.

Dr (Prof) R.S.ThakuR, Professor Department of mathematics, Bioinformatics & Computer Application MANIT Bhopal (MP) 9826241996.ramthakur2000@gmail.com.