Asian Journal of Computer Science and Technology
ISSN: 2249-0701 (P) Vol.7 No.S1, 2018, pp.96-100
© The Research Publication, www.trp.org.in
DOI: https://doi.org/10.51983/ajcst-2018.7.S1.1796

An Empirical Review on Data Feature Selection and Big Data Clustering

Venkata Rao Maddumala[1], R. Arunkumar[2] and S. Arivalagan[3]
[1]Department of Computer Science and Engineering, Annamalai University, Tamil Nadu, India
[2&3]Department of Information Technology,
Vignan's Nirula Institute of Technology of Science for Women, Andhra Pradesh, India
E-Mail: venkatresearchau@gmail.com

(Received 3 October 2018; Revised 18 October 2018; Accepted 31 October 2018; Available online 8 November 2018)

Abstract - **With the fast advancement of the Big Data, Big Data innovations have risen as a key data investigation apparatus, in which, feature extraction and data bunching calculations are considered as a basic part for data examination. Nonetheless, there has been constrained research that tends to the difficulties crosswise over Big Data and along these lines proposing an exploration motivation is vital to illuminate the examination challenges for bunching Big Data. By handling this particular viewpoint - grouping calculation in Big Data, this paper looks at on Big Data advancements, identified with feature determination and data bunching calculations and conceivable uses. In view of our survey, this paper distinguishes an arrangement of research difficulties that can be utilized as an exploration plan for the Big Data bunching research. This exploration plan goes for distinguishing and crossing over the examination holes between Big Data feature choice and grouping calculations.**

Keywords: **Big Data, Clustering, Feature Selection**

I. INTRODUCTION

Enormous Data is defended as indicated by three central components, which are volume (size of information), assortment (diverse sorts of information from a few sources) and speed (information gathered continuously). In addition, other research work acquainted extra attributes with the 3V's model, for example, that displayed further angles: esteem (advantages to different modern and scholastic fields), veracity, legitimacy (adjust preparing of the information), changeability, consistency (dormancy information transmission between the source and goal), virility (speed of the information send and gets from different sources) and representation (understanding of information is more concerned and ID of the most applicable data for the clients) [1-4]. Notwithstanding the presence of extra qualities of Big Data, the 3V's model sets the premise of the Big Data idea. The combination of Big Data and IoT advances has made open doors for the improvement of administrations for shrewd conditions like savvy urban areas. There have been in this manner a few Big Data advances accessible to help the handling of vast volume of IoT information have risen as a need to process the information gathered from various sources in the savvy condition. Be that as it may, the headway of IoT is progressively creating immense sum and distinctive kinds of information, particularly after the presence of the rising 5G. In the meantime, Big Data and its advancements have opened new opportunities for enterprises and scholastics to grow new IoT arrangements. Accordingly , the combination of Big Data and IoT, and in addition the very powerful advancement of the two areas, makes new research difficulties, which have so far not been perceived and tended to by the exploration network. This paper handles a particular and vital part of the Big Data, grouping calculations in Big Data, as bunching is a basic task for Big Data preparing and investigation. We have surveyed the points of interest and disservices of grouping calculations, which show that bunching is one of the key variables to supply the combination of Big Data, distributed computing, portable condition and IoT innovations [5-7]. The commitments of the paper are twofold: we have inspected the grouping calculations in Big Data and showed how bunching calculations in Big Data can be utilized in IoT. In view of the audit, we have proposed an arrangement of research difficulties to clear up the exploration holes between Big Data and IoT.

A. Data Mining Algorithm for Map-Reduce Solution

As we made reference to in the past areas, the majority of the conventional information mining calculations are not intended for parallel figuring; along these lines, they are not especially helpful for the huge information mining. A few ongoing examinations have endeavored to adjust the customary information mining calculations to make them pertinent to Hadoop-based stages. For whatever length of time that porting the information mining calculations to Hadoop is inescapable, making the information mining calculations chip away at a guide lessen engineering is the main exceptionally activity to apply customary information mining strategies to huge information investigation. Tragically, very few examinations endeavored to make the information mining and delicate figuring calculations take a shot at Hadoop on the grounds that few distinct foundations are expected to create and plan such calculations. For example, the scientist and his or her examination amass need the foundation in information mining and Hadoop in order to create and structure such calculations. Another open issue is that most information digging calculations are intended for incorporated figuring; that is, they can just work on every one of the information in the meantime. Consequently, how to make them chip away at a parallel registering framework is likewise a troublesome work. Fortunately a few investigations have effectively connected the customary information mining calculations to the guide decrease design. These outcomes suggest that it is conceivable to do as such. As per our perception, in spite of the fact that the customary mining or delicate processing calculations can be utilized to enable us to break down the information in enormous information investigation, tragically, as of recently, very few examinations are centered on it. As a result, it is an essential open issue in enormous information examination.

II. FEATURE EXTRACTION AND FEATURE SELECTION

There are a few methodologies in the writing to deal with feature mining and feature choice while working with transient information. Following the Singular Spectrum Analysis, referred to likewise as the Caterpillar technique [8], a period arrangement for a window length L, is spoken to by an arrangement of multi-dimensional vectors appeared as $X \times Y$ frameworks. Along these lines, such a Y X X lattice X relates to the time $X + Y - 1$, to signify M ongoing qualities in X. At the point when another perception (i.e. an occasion) arrives, lattice X is changed to another one, where the primary record is discarded while another record/line is included as the last line of the network. This methodology performs productively on account of having a fairly modest number of measurements; as we continue with the examination of multi-dimensional heterogeneous clear cut information, it neglects to display the time-arrangement viably. The creators in [9] contend that application payload, including all out heterogeneous information, is portrayed by successive information which is not relevant for learning techniques that work in metric spaces.

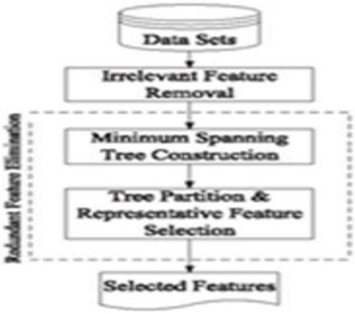
Data Sets
Irrelevant Feature Removal
Minimum Spanning Tree Construction
Tree Partition & Representative Feature Selection
Redundant Feature Elimination
Selected Features

Fig. 1 Feature selection architecture

They contend that features can be extricated either as parse trees speaking to punctuation features or as consecutive features. In this paper we make utilization of both of these sorts of features. In Apache Spark all out qualities should be changed into the Double sort or into Vectors of quantities of the Double kind, after feature extraction. Moreover, in this work, we make utilization of the accessible feature extractors, for example, the TF-IDF, Word2Vec, Count Vectorizer (to make the information vocabulary of the considerable number of features over the records, following the methodology displayed in [10]). For the necessities of our work, we have expanded the usefulness of the NLTK bundle [11] to extricate the features from complex structures. In Section III, we present a methodology for feature extraction using Spark, on account of working with staggered complex contributions from various sources. As indicated by [12], the undertaking of features choice which portray accumulated streams is a key issue to any interruption recognition framework, while it winds up basic and testing on account of unsupervised identification, because of the absence of any data of which may be the most significant arrangement of features.

A methodology for feature choice in time arrangement utilizing independent framework stream totals is introduced in [13], which diminishes the quantity of streams to be examined in the subsequent stages of information examination by embracing a moving windowing strategy. Time is separated into interims (i.e. eras over which a solitary factual model will be allocated amid the preparation stage), ages (i.e. eras over which tests are gathered), and accumulation periods (i.e. eras over which a metric of intrigue is amassed). Their methodology is exceptionally centered on IP stream conglomeration, without considering some other framework substance or system benefit. In [14], with the end goal to separate and select the correct features for characterizing movement records for framework substances, they propose the utilization of registered social features over an interim of time, e.g. a day-hour window alongside totaled movement records through covering time scopes. Along these lines, they figure out how to accomplish constant necessities, as insights are as of now ascertained for the past time space: for instance, the calculation of features, for example, any checks, midpoints, standard deviations, amassing pointers, social features at the crossing point of two elements, or worldly practices between at least two occasions in time utilizing timestamps, over the past 24 hours requires the recovery and conglomeration of 82 records.

III. CLUSTERING ALGORITHMS IN BIG DATA

Clustering algorithms have developed as a preprocessing instrument to learn and break down the Big Data [15]. The objective of clustering calculations is to aggregate the information in a similar group dependent on certain likeness measurements. There as of now exists various bunching calculations, and in addition thinks about that examine their focal points and downsides. As [16] showed, grouping calculations are right now advancing to address diverse Big Data difficulties. This segment in this manner surveys distinctive bunching calculations for Big Data, which can be utilized in IoT. Albeit a few investigations likewise proposed promising adaptable parallel programming models that can bolster parallel grouping calculations for taking care of Big Data, auditing the parallel bunching calculations dependent on MapReduce isn't in the extent of this paper. Bunching is a fundamental information mining utilized as a Big Data investigation strategy. The guideline of this system is to make gatherings or subsets that contain the items with comparable trademark highlights. Thusly, the bunch investigation makes information control straightforward by discovering structure in information and arranging each question as indicated by its tendency. In addition, it is isolated into two classifications: single machine grouping strategies, which utilize assets of only one single machine, and different machine bunching procedures, which keep running in a few machines and approach more assets. In this segment we endeavor to sort the larger part of accessible grouping calculations as indicated by their pertinence in Big Data as pursues: Hierarchical calculation: The objective of various leveled bunching is to assemble a progressive tree to demonstrate the connection of groups in two unique behavior, which are agglomerative strategy and disruptive technique. Agglomerative technique begins with one-point (singleton) groups and recursively includes at least two proper bunches until the point when it accomplishes a K number of bunches. Then again, disruptive strategy isolates the information to a solitary group, which contains all information objects, into littler proper bunches until the point when a halting basis is accomplished.

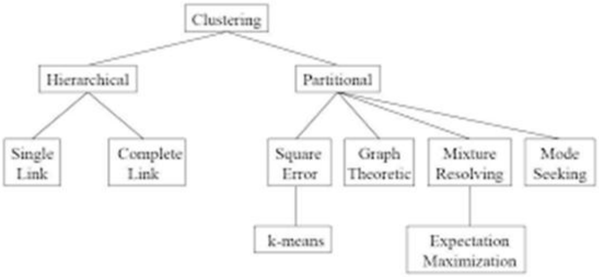

Fig. 2 Different clustering algorithms

A. Partitional Calculation: Unlike the various leveled grouping calculations that force a progressive structure, the partitional calculations discover every one of the bunches all the while as an underlying parcel of the information. At that point the items are doled out to the comparative group focus dependent on particular criteria.

B. Thickness Based Calculation: The principle of utilizing these systems is to find groups of various shapes and sizes from substantial datasets, where each bunch is spoken to by a maximal arrangement of thickness associated objects, which are part founded on the locale of thickness, network and limit. Because of high computational unpredictability, this sort of strategies can enhance promote the correspondence cost.

C. Centroid-Based Clustering Calculation: The general thought of this strategy is that each group is spoken to by a question or, in other words midway situated in a bunch. Besides, the centroid-based calculation lessens all examinations among articles and bunches into basic correlations among items and the medoids of the groups. Single-linkage progressive calculation: all in all, a solitary linkage calculation is one of a few strategies for various leveled bunching; it expects to lessen computational multifaceted nature by joining two diverse groups dependent on the separation between the nearest two articles [17]. Be that as it may, it can deliver anchoring impact if the bunches are considerably more distant from one another than to objects of other.

D. Network Based Calculation: The information space is parceled into a limited number of matrices and a bunch is spoken to by an area that has a maximal arrangement of thickness focuses [18]. The quantity of matrices is littler than the quantity of examples. Subsequently, in the parceling phase of this sort of strategies, the networks could create great outcomes as far as grouping time. Similitude based grouping calculation: The principle thought of this down to earth system is to quantify the likeness of two questions and decide whether they are comparative or different [19-20]. In light of the level of likeness, comparative articles are put away in a similar bunch and disparate items are in various groups. Be that as it may, this calculation is unequipped for managing huge information occasions.

E. Co-Bunching Calculation: Unlike to the customary grouping calculations that contain a comparative subset of the lines over all segments, co-grouping calculation corresponds the subsets of lines with just a subset of its segments [21]. Notwithstanding, it isn't reasonable to apply on vast informational index. Inside the best in class, works exist that study grouping calculations to decide the best performing for Big Data [22]. K-implies is a standout amongst the most utilized grouping calculations in Big Data [23-24]. It is an apportioned bunching calculation that accepts K as starting group focuses (input parameter). Next, it parcels an arrangement of n objects into K bunches. At that point it decides group similitude or bunch focus as per the mean estimation of the items in the group. In view of the separation between the question and the bunch focus, it relegates each protest the group to which it is the most comparative. At long last, it figures the new mean for each group. Cop-k-implies is an adjusted adaptation of k-implies, where two pair wise of limitations, to be specific, Must-interface (ML) and Cannot-connect (CL), are added to maintain a strategic distance from computational conditions between Mappers. Therefore, the task of items to bunch is organization delicate. PSO [25] takes care of the affectability issue of k-implies on introductory group focus by executing three MapReduce occupations, where the principal work creates new molecule centroids, the second employment utilizes the wellness capacity to assess the new molecule centroids, which are produced in the main module. At long last, the third employment combines the wellness esteems that are the yields of the first and second modules. PAM is another grouping approach that has a place with centroid-based bunching [26-27]. It picks k arbitrary protests as the underlying medoids. At that point it ascertains the separation between each question and k medoids with the end goal to relegate each protest the group with the nearest medoid. Rather than PAM, CLARA, or, in other words of PAM calculation, centers to group little arbitrary subsets of the dataset. In this way, the entire emphasis is diminished into two MapReduce employments: The main occupation computes arbitrary subsets and the second estimates the quality. Thus, it accomplishes insignificant employment dormancy in light of the fact that the information is just stacked twice. On account of the likeness that estimates the rationality of the items and chooses naturally the comparable subsets, co-bunching calculation dependent on MapReduce has demonstrated its effectiveness and unwavering quality in numerous areas, for example, the enhancement of disease subtype recognizable proof [28]. Numerous works center around running DBSCAN (thickness based spatial grouping of uses with commotion) calculation in MapReduce, for example [29]. The general thought of DBSCAN is to conquer the impact of commotion and find bunches of subjective shape. To do that, the articles are part founded on the locale of thickness, network and limit. Next, a group is shaped by a maximal arrangement of thickness associated objects that are maximal thickness reachable. At that point the calculation utilizes a pioneer thickness based grouping calculation to recognize self-assertively molded bunches. Be that as it may, a ton of I/O overhead is delivered because of the need to recognize each protest decide if it is the center question. Also, it performs ineffectively if the bunches having distinctive densities. Output [30-31] is an augmentation of DBSCAN approach for expansive systems. The upside of this calculation is to distinguish the initiated vertices as new individuals from the bunch to deal with huge systems with a huge number of vertices. STING (Statistical Information Grid-based technique) is one of the delegate grouping calculations dependent on matrix, which bunches spatial information. Like the grouping properties of record structures, the spatial zone is separated freely into rectangular network cells at various levels and every cell at level is apportioned into k number of cells at the following level, which frames a progressive structure that procedures the insights data put away in framework units. To accomplish more inclination in dispersed condition, STING calculation is executed utilizing MapReduce and Hadoop. Notwithstanding, the grouping calculations dependent on framework are significantly touchy to the high granularity of matrix, which can diminish the nature of bunching and additionally the grouping precision. Because of the upsides of matrix worldview, the well-known single linkage various leveled grouping calculation (SLINK) is joined with the framework to create GridSLINK that plans to decrease the quantity of separation counts required by SLINK. Not at all like the conventional strategies that consider the similitude esteems from examples to k focuses, otherworldly calculation can identify complex nonlinear structures, and select bunches dependent on pair wise likenesses of information occasions. Be that as it may, it requires significant memory and computational time when the extent of information examples is substantial. Another bunching calculation is called CURE [32-34], or, in other words grouping algorithm.

Given the nature of the Big Data, we have proposed the following research questions.

1. How to adequately choose the feature determination calculations for Big Data?
2. How to adequately choose the bunching calculations for Big Data?
3. How to powerfully choose the most reasonable calculation to bunch the Big Data in an opportune way?
4. How to group the diverse kinds of Big Data that speak to the equivalent or comparable element/occasion?
5. How to pick the best possible Big Data advancements, for example, MapReduce structures to play out the grouping calculation for Big Data?

F. Noise, Outliers, Incomplete and Inconsistent Data: Albeit huge information examination is another age for information investigation, in light of the fact that few arrangements embrace established approaches to dissect the information on enormous information investigation, the open issues of conventional information mining calculations additionally exist in these new frameworks. The open issues of commotion, anomalies, deficient, and conflicting information in customary information mining calculations will likewise show up in huge information mining calculations. More fragmented and conflicting information will effectively show up in light of the fact that the information are caught by or created from various sensors and frameworks. The effect of commotion, anomalies, inadequate and conflicting information will be broadened for enormous information investigation. In this way, how to relieve the effect will be the open issues for enormous information investigation.

IV. BOTTLENECKS ON DATA MINING ALGORITHM

The majority of the information mining calculations in enormous information investigation will be intended for parallel figuring. In any case, when information mining calculations are structured or changed for parallel registering, it is the data trade between various information mining methodology that may acquire bottlenecks. One of them is the synchronization issue on the grounds that distinctive mining techniques will complete their employments at various occasions despite the fact that they utilize a similar mining calculation to chip away at a similar measure of information. In this way, a portion of the mining techniques should hold up until the point when the others completed their employments. This circumstance may happen in light of the fact that the stacking of various PC hubs might be distinctive amid the information mining procedure, or it might happen in light of the fact that the assembly speeds are diverse for similar information mining calculation. The bottlenecks of information mining calculations will turn into an open issue for the huge information investigation which discloses that we have to consider this issue when we create and plan another information digging calculation for huge information examination.

V. CONCLUSION

In this article, we have led a review on Big Data innovations and grouping calculations, in which we have determined the upsides and downsides of every calculation in the Big Data setting. We have additionally related our audit to the exploration of IoT and examined the relations between Big Data, bunching calculations and IoT. In view of the audit, we have anticipated a course of action of research difficulties that location the rising exploration points and research questions on the feature selection and data clustering. The exploration difficulties can be considered as an examination motivation to manage the future research crosswise over Big Data people group. In particular, this paper has stressed the significance of bunching calculations in Big Data and brings consideration and conceivable uses of the Big Data grouping calculations.

VI. FUTURE WORK

As impending work, we intend to additionally examine the associations between the systems of Big Data feature determination and grouping. We will break down which Big Data innovations can be successfully utilized in which setting.

REFERENCES

[1] Al-Madi, Nailah, Ibrahim Aljarah, and Simone A. Ludwig, "Parallel glowworm swarm optimization clustering algorithm based on MapReduce", *IEEE Symposium on Swarm Intelligence, 2014.*

[2] Amini, Amineh, Teh Ying Wah, and Hadi Saboohi, "On density-based data streams clustering algorithms: a survey", *Journal of Computer Science and Technology*, Vol. 29, No.1, pp. 116-141, 2014.

[3] A. Akbar, F. Carrez, K. Moessner, J. Sancho and J. Rico, "Context-aware stream processing for distributed IoT applications In Internet of Things (WF-IoT)", *2015 IEEE 2nd World Forum*, pp. 663- 668, Dec. 2015.

[4] Ahmed, Ejaz, and Mubashir Husain Rehmani, "Mobile edge computing: opportunities, solutions, and challenges", pp. 59-63, 2017.

[5] Pavel Berkhin, "A survey of clustering data mining techniques", Grouping multidimensional data. *Springer Berlin Heidelberg*, pp. 25-71, 2006.

[6] Chen, Yong, Hong Chen, Anjee Gorkhali, Yang Lu, Yiqian Ma, and Ling Li, "Big Data analytics and Big Data science: a survey", *Journal of Management Analytics 3*, Vol. 1, pp. 1- 42, 2016.

[7] Da Xu, Li, Wu He, and Shancang Li, "Internet of things in industries: A survey", *IEEE Transactions on industrial informatics*, Vol. 10, No. 4, pp. 2233-2243, 2014.

[8] D'Urso Pierpaolo, Riccardo Massari, Livia De Giovanni, and Carmela Cappelli, "Exponential distance-based fuzzy clustering for interval-valued data", *Fuzzy Optimization and Decision Making* Vol. 16, No.1, pp. 51-70, 2017.

[9] Sanjit Kumar Dash, Debi Prasad Mishra, Ranjita Mishra, and Sweta Dash, "Privacy preserving K-Medoids clustering: an approach towards securing data in Mobile cloud architecture" *2nd International Conference on Computational Science, Engineering and Information Technology*, pp. 439-443, ACM, 2012.

[10] Anind K Dey, "Understanding and using context", *Personal and ubiquitous computing*, Vol. 5, No.1, pp. 4-7, 2001.

[11] El Naqa, Issam and Martin J. Murphy, "What Is Machine Learning?" *Machine Learning in Radiation Oncology*. Springer International Publishing, pp. 3-11, 2015.

[12] Fahad, Adil, Najlaa Alshatri, Zahir Tari, Abdullah Alamri, Ibrahim Khalil, Albert Y. Zomaya, Sebti Foufou, and Abdelaziz Bouras, "A survey of clustering algorithms for Big Data: Taxonomy and empirical analysis", *IEEE transactions on emerging topics in computing*, Vol. 2, No. 3, pp. 267-279, 2014.

[13] Fredj, Sameh Ben, Mathieu Boussard, Daniel Kofman, and Ludovic Noirie, "A scalable IoT service search based on clustering and aggregation", *In Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things*, pp. 403-410, 2013.

[14] Kun Guo, Wenzhong Guo, Yuzhong Chen, Qirong Qiu, and Qishan Zhang, "Community discovery by propagating local and global information based on the MapReduce model", *Information Sciences*, Vol. 323, pp. 73-93, 2015.

[15] Poonam Goyal, Sonal Kumari, Sumit Sharma, Dhruv Kumar, Vivek Kishore, Sundar Balasubramaniam, and Navneet Goyal, "A Fast, Scalable SLINK Algorithm for Commodity Cluster Computing Exploiting Spatial Locality", *In High Performance Computing and Communications*; *IEEE 14th International Conference on Smart City*, 2016.

[16] Timothy C. Havens, James C. Bezdek, and Marimuthu Palaniswami, "Scalable single linkage hierarchical clustering for Big Data", *Intelligent Sensors, Sensor Networks and Information Processing, 2013 IEEE Eighth International Conference on. IEEE*, 2013.

[17] Hossain, M. Shamim, Changsheng Xu, Ying Li, Al-Sakib Khan Pathan, Josu Bilbao, Wenjun Zeng, and Abdulmotaleb El Saddik, "Impact of Next-Generation Mobile Technologies on IoT-Cloud Convergence", *IEEE Communications Magazine*, Vol. 55, No. 1, pp. 18-19, 2017.

[18] Jiang, Dajie, and Guangyi Liu, "An Overview of 5G Requirements", *5G Mobile Communications. Springer International Publishing*, pp. 3-26, 2017.

[19] Kitchin, Rob, "Big Data—Hype or revolution", *The SAGE handbook of social media research methods*, 2017.

[20] Liu, Yiyi, Quanquan Gu, Jack P. Hou, Jiawei Han, and Jian Ma, "A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression", *BMC bioinformatics*, Vol. 15, No. 1, pp. 37, 2014.

[21] Lin, Chao, Yan Yang, and Tonny Rutayisire, "A parallel Cop-K means clustering algorithm based on MapReduce framework", *Knowledge Engineering and Management*, pp. 93-102, 2011.

[22] Li, Yan, Hong Liu, Guang-peng Liu, Liang Li, Philip Moore, and Bin Hu, "A grouping method based on grid density and relationship for crowd evacuation simulation", *Physical A: Statistical Mechanics and its Applications*, 2017.

[23] Manogaran, Gunasekaran, Chandu Thota, Daphne Lopez, V. Vijayakumar, Kaja M. Abbas, and Revathi Sundarsekar, "Big Data Knowledge System in Healthcare", *In Internet of Things and Big Data Technologies for Next Generation Healthcare*, pp. 133- 157, Springer International Publishing, 2017.

[24] Mavromoustakis, Constandinos X., George Mastorakis, and Jordi Mongay Batalla, "Internet of Things (IoT) in 5G Mobile Technologies", *Modeling and Optimization in Science and Technologies*, 2016

[25] Mohebi, Amin, Saeed Aghabozorgi, Teh Ying Wah, Tutut Herawan, and Ramin Yahyapour, "Iterative Big Data clustering algorithms: a review", *Software: Practice and Experience*, Vol. 46, No. 1, pp. 107-129, 2016.

[26] Nguyen, Cuong Duc, Dung Tien Nguyen, and Van-Hau Pham, "Parallel two-phase K-means", *International Conference on Computational Science and Its Applications*. Springer Berlin Heidelberg, 2013.

[27] Ng, Raymond T., and Jiawei Han, "CLARANS: A method for clustering objects for spatial data mining", *IEEE transactions on knowledge and data engineering*, Vol. 14, No. 5, pp. 1003-1016, 2002.

[28] Pandove, Divya, and Shivani Goel, "A comprehensive study on clustering approaches for Big Data mining", *Electronics and Communication Systems (ICECS), 2015 2nd International Conference on IEEE, 2015*.

[29] Rafailidis, D., E. Constantinou and Y. Manolopoulos, "Landmark selection for spectral clustering based on Weighted Page Rank", *Future Generation Computer Systems*, Vol. 68, pp. 465 - 472, 2017.

[30] Shirkhorshidi, Ali Seyed, Saeed Aghabozorgi, Teh Ying Wah, and Tutut Herawan, "Big Data clustering: a review", *In International Conference on Computational Science and Its Applications*, pp. 707-720, 2014.

[31] Srirama, Satish Narayana, Pelle Jakovits, and Eero Vainikko, "Adapting scientific computing problems to clouds using MapReduce", *Future Generation Computer Systems*, Vol. 28, No. 1, pp. 184-192, 2012.

[32] Sreenivasulu, G., S. Viswanadha Raju, and N. Sambasiva Rao, "Review of Clustering Techniques", *International Conference on Data Engineering and Communication Technology*. Springer Singapore, 2017.

[33] Van Kranenburg, Rob. The Internet of Things: A critique of ambient technology and the all-seeing network of RFID. *Institute of Network Cultures*, 2008.

[34] Xu, Lina, Rem Collier, and Gregory MP O'Hare, "A survey of clustering techniques in WSNs and consideration of the challenges of applying such to 5g iot scenarios", *IEEE Internet of Things Journal*, Vol. 4, No. 5, pp. 1229-1249, 2017.