Exploring Big Data Clustering Algorithms for Internet of Things Applications

Hind Bangui[1,2,3], Mouzhi Ge[1,2] and Barbora Buhnova[1,2]
[1]*Institute of Computer Science, Masaryk University, Brno, Czech Republic*
[2]*Faculty of Informatics, Masaryk University, Brno, Czech Republic*
[3]*FSTG, Cadi Ayyad University, Marrakesh, Morocco*

Keywords: Big Data, Internet of Things, Clustering Algorithm, Machine Learning, Mobile Networks.

Abstract: With the rapid development of the Big Data and Internet of Things (IoT), Big Data technologies have emerged as a key data analytics tool in IoT, in which, data clustering algorithms are considered as an essential component for data analysis. However, there has been limited research that addresses the challenges across Big Data and IoT and thus proposing a research agenda is important to clarify the research challenges for clustering Big Data in the context of IoT. By tackling this specific aspect - clustering algorithm in Big Data, this paper examines on Big Data technologies, related data clustering algorithms and possible usages in IoT. Based on our review, this paper identifies a set of research challenges that can be used as a research agenda for the Big Data clustering research. This research agenda aims at identifying and bridging the research gaps between Big Data clustering algorithms and IoT.

1 INTRODUCTION

Internet of Things (IoT) is one of the most promising technologies in the current epoch. IoT is characterized by using smart and self-configuring objects that can interact with each other via global network infrastructure. Therefore, these interactions between large amounts of heterogeneous objects represent IoT as a disruptive technology that enables ubiquitous and pervasive computing applications (Van Kranenburg 2008). Accordingly, a wide range of industrial IoT applications (Da et al., 2014) have been developed and deployed in different domains such as transportation systems, agriculture, food processing industry, health monitoring systems, environmental monitoring, and security surveillance.

Since IoT connects the sensors and other devices to the Internet, it plays an important role to support the development of smart services. In other words, the dynamic things collect different kinds of information from the real-world environment. Thus, the extracted relevant information from IoT data can be used to improve and enrich our daily life with context-aware applications, which can for example display contents related to the current situation of users (Dey, 2001). It can thus identify the data relevant to an object based on the object's contextual information. IoT is an important source of contextual data with a large vomue and fast velocity that is considered as typical characristics of Big Data.

Big Data is definded according to three fundamental elements, which are volume (size of data), variety (different types of data from several sources) and velocity (data collected in real time). Moreover, other research work introduced additional characteristics to the 3V's model such as (Manogaran et al., 2017) that presented further aspects: value (benefits to various industrial and academic fields), veracity (uncertainty of data), validity (correct processing of the data), variability (context of data), viscosity (latency data transmission between the source and destination), virality (speed of the data send and receives from various sources) and visualization (interpretation of data is more concerned and identification of the most relevant information for the users). Despite the existence of additional characteristics of Big Data, the 3V's model sets the basis of the Big Data concept (Kitchin, 2017).

The fusion of Big Data and IoT technologies has created opportunities for the development of services for smart environments like smart cities. There have been thus several Big Data technologies available to support the processing of large volume of IoT data

Bangui, H., Ge, M. and Buhnova, B.
Exploring Big Data Clustering Algorithms for Internet of Things Applications.
DOI: 10.5220/0006773402690276
In *Proceedings of the 3rd International Conference on Internet of Things, Big Data and Security (IoTBDS 2018)*, pages 269-276
ISBN: 978-989-758-296-7

such as Big Data analytics (Chen et al., 2016), which have emerged as a need to process the data collected from different sources in the smart environment. However, the advancement of IoT is increasingly producing vast amount and different types of data, especially after the appearance of the emerging 5G (Mavromoustakis et al., 2016). At the same time, Big Data and its technologies have opened new oppotunities for industries and academics to develop new IoT solutions. Therefore , the fusion of Big Data and IoT, as well as the highly dynamic evolution of the two domains, creates new research challenges, which have so far not been recognized and addressed by the research community.

This paper tackles a specific and important aspect of the Big Data, clustering algorithms in Big Data, as clustering is a critical operation for Big Data processing and analytics. We have reviewed the advantages and disadvantages of clustering algorithms, which indicate that clustering is one of the key factors to supply the fusion of Big Data, cloud computing, mobile environment and IoT technologies. The contributions of the paper are two-fold: we have reviewed the clustering algorithms in Big Data and illustrated how clustering algorithms in Big Data can be used in IoT. Based on the review, we have proposed a set of research challenges to clarify the research gaps between Big Data and IoT.

The remainder of the paper is organized as follows. Section 2 carries out a literature review on clustering algorithms in Big Data, which includes algorithm characteristics and classification. Based on the review, Section 3 presents the major challenges and opportunities related to the fusion of IoT, cloud computing, mobile environment and Big Data technologies in the 5G networks. Finally, section 4 concludes the work and outlines future research.

2 CLUSTERING ALGORITHMS IN BIG DATA

Clustering algorithms have emerged as a pre-processing tool to learn and analyze the Big Data (Fahad et al., 2014). The goal of clustering algorithms is to group the data in the same cluster based on certain similarity metrics. There already exists a number of clustering algorithms, as well as studies that discuss their advantages and drawbacks (Shirkhorshidi et al., 2014). As (Fahad et al., 2014) indicated, clustering algorithms are currently evolving to meet different Big Data challenges. This section therefore reviews different clustering algorithms for Big Data, which can possibly be used in IoT. Although some studies also proposed promising flexible parallel programming models that can support parallel clustering algorithms for handling Big Data (Mohebi et al., 2016), reviewing the parallel clustering algorithms based on MapReduce is not in the scope of this paper.

Clustering is an essential data mining used as a Big Data analytics method. The principle of this technique is to create groups or subsets that contain the objects with similar characteristic features (Havens et al., 2013). Consequently, the cluster analysis makes data manipulation simple by finding structure in data and classifying each object according to its nature. Besides, it is divided into two categories: single machine clustering techniques, which use resources of just one single machine, and multiple machine clustering techniques, which run in several machines and have access to more resources. In this section we try to categorize the majority of available clustering algorithms according to their applicability in Big Data as follows:

Hierarchical algorithm: The goal of hierarchical clustering is to build a hierarchical tree to show the relation of clusters in two different manners, which are agglomerative method and divisive method (Pandove and Goel, 2015). Agglomerative method starts with one-point (singleton) clusters and recursively adds two or more appropriate clusters until it achieves a K number of clusters. On the other hand, divisive method divides the data to a single cluster, which contains all data objects, into smaller appropriate clusters until a stopping criterion is achieved.

Partitional algorithm: Unlike the hierarchical clustering algorithms that impose a hierarchical structure, the partitional algorithms find all the clusters simultaneously as an initial partition of the data. Then the objects are assigned to the similar cluster center based on specific criteria (Nguyen et al. 2013, Lin et al. 2011, Al-Madi et al. 2014).

Density-based algorithm: The main of using these techniques is to discover clusters of different shapes and sizes from large datasets, where each cluster is represented by a maximal set of density-connected objects, which are split based on the region of density, connectivity and boundary (Amini et al. 2014, Guo et al. 2015). Due to high computational complexity, this kind of methods is able to improve further the communication cost.

Centroid-based clustering algorithm: The general idea of this technique is that each cluster is represented by an object (medoid), which is the most centrally located in a cluster (Srirama et al. 2012, Ng

and Han 2002). Moreover, the centroid-based algorithms reduce all comparisons between objects and clusters into simple comparisons between objects and the medoids of the clusters.

Single-linkage hierarchical algorithm: In general, a single-linkage algorithm is one of several methods of hierarchical clustering, it aims to reduce computational complexity by combining two different cluster based on the distance between the closest two objects (Goyal et al. 2016, Rafailidis and Manolopoulos 2017). However, it can produce chaining effect if the clusters are much farther from each other than to objects of other.

Grid-based algorithm: The data space is partitioned into a finite number of grids and a cluster is represented by a region that has a maximal set of density points (Wang et al., 1997). The number of grids is smaller than the number of instances. Consequently, in the partitioning stage of this kind of methods, the grids could produce good results in terms of clustering time.

Similarity-based clustering algorithm: The main idea of this practical technique is to measure the similarity of two objects and determine if they are similar or dissimilar (Fahad et al. 2014, Shirkhorshidi et al. 2014). Based on the degree of similarity, similar objects are stored in the same cluster and dissimilar objects are in different clusters. However, this algorithm is incapable of dealing with massive data instances.

Co-clustering algorithm: Unlike to the traditional clustering algorithms that contain a similar subset of the rows across all columns, co-clustering algorithm correlates the subsets of rows with only a subset of its columns (Liu et al. 2014). However, it is not practical to apply on large data set.

Within the state of the art, works exist that survey clustering algorithms to determine the best performing for Big Data (Fahad et al., 2014, Berkhin 2006). K-means is one of the most used clustering algorithms in Big Data (Nguyen et al., 2013). It is a partitioned clustering algorithm that takes K as initial cluster centers (input parameter). Next, it partitions a set of n objects into K clusters. Then it determines cluster similarity or cluster center according to the mean value of the objects in the cluster. Based on the distance between the object and the cluster center, it assigns each object to the cluster to which it is the most similar. Finally, it calculates the new mean for each cluster. Cop-k-means (Lin et al., 2011) is a modified version of k-means, where two pairwise of constraints, namely, Must-link (ML) and Cannot-link (CL), are added to avoid computational dependencies between Mappers. Consequently, the assignment of objects to clusters is order-sensitive. PSO (Al-Madi et al. 2014) solves the sensitivity problem of k-means on initial cluster center by executing three MapReduce jobs, where the first job generates new particle centroids, the second job uses the fitness function to evaluate the new particle centroids, which are generated in the first module. Finally, the third job merges the fitness values that are the outputs of the first and second modules.

PAM (Partitioning Around Medoids) is another clustering approach that belongs to centroid-based clustering (Srirama et al., 2012). It chooses k random objects as the initial medoids. Then it calculates the distance between each object (non-medoid) and k medoids in order to assign each object to the cluster with the closest medoid. In contrast to PAM, CLARA (Clustering Large Application) (Ng and Han 2002, Srirama et al. 2012), which is an improvement of PAM algorithm, focuses to cluster small random subsets of the dataset. So, the whole iteration is reduced into two MapReduce jobs: The first job calculates random subsets and the second measures the quality. As a result, it achieves minimal job latency because the input data is only loaded twice. Thanks to the similarity that measures the coherence of the objects and selects automatically the similar subsets, co-clustering algorithm based on MapReduce has proved its efficiency and reliability in many domains such as the improvement of cancer subtype identification (Liu et al., 2014). Many works focus on running DBSCAN (density-based spatial clustering of applications with noise) algorithm in MapReduce such as (Amini et al., 2014). The general idea of DBSCAN is to overcome the effect of noise and discover clusters of arbitrary shape. To do that, the objects are split based on the region of density, connectivity and boundary. Next, a cluster is formed by a maximal set of density-connected objects that are maximal density reachable. Then the algorithm uses a pioneer density based clustering algorithm to detect arbitrarily shaped clusters. However, a lot of I/O overhead is produced due to the need to detect each object to determine whether it is the core object. Besides, it performs poorly if the clusters having different densities. SCAN (Structural Clustering Algorithm for Networks) (Guo et al., 2015) is an extension of DBSCAN approach for large networks. The advantage of this algorithm is to identify the activated vertices as new members of the cluster to handle big networks with millions of vertices.

STING (Statistical Information Grid-based method) (Wang et al., 1997) is one of the representative clustering algorithms based on grid, which clusters spatial data. Similar to the clustering

properties of index structures, the spatial area is divided independently into rectangular grid cells at different levels and each cell at level is partitioned into k number of cells at the next level, which forms a hierarchical structure that processes the statistics information stored in grid units. To achieve more preference in distributed environment, STING algorithm is implemented using MapReduce and Hadoop (Li et al., 2017). However, the clustering algorithms based on grid are greatly sensitive to the high granularity of grid, which can decrease the quality of clustering as well as the clustering accuracy. Due to the advantages of grid paradigm, the popular single linkage hierarchical clustering algorithm (SLINK) is combined with the grid to produce GridSLINK (Goyal et al., 2016) that aims to reduce the number of distance calculations required by SLINK.

Unlike the traditional methods that consider the similarity values from instances to k centers, spectral algorithm (Rafailidis and Manolopoulos, 2017) is able to detect complex nonlinear structures, and select clusters based on pairwise similarities of data instances. However, it requires considerable memory and computational time when the size of a data instances is large.

Another clustering algorithm is called CURE (Clustering Using REpresentative) (Pandove and Goel, 2015), which is a hierarchical clustering algorithm. The general idea around this algorithm is to present each cluster by fixed number of well scattered data points in order to determine general shapes. Moreover, it aims at reducing the dimension of the input data matrix by transforming it into a lower dimensional space. As a result, the effect of outliers and noise in this reduced space are reduced. To improve further its efficiency for Big Data, the algorithm is implemented over distributed environment using MapReduce and Hadoop.

Despite the advantages of the clustering algorithms mentioned above, the extraction of useful information out of massive amount of data and process them in less time has become a crucial operation for the existing clustering algorithms, which use non-iterative parallel programming models that require re-clustering each time new object is generated. Yet, we need to face the challenges associated with map and reducer programming paradigm notably MapReduce, which is considered as the most programing model adapted from clustering algorithms (Table I) and used generally in Big Data architecture.

Clustering algorithm is one of the important Big Data analytics methods that are used to extract useful information from data (Sreenivasulu et al., 2017). Moreover, the extraction of data is the most crucial process in Big Data. The main goal of Big Data is to produce efficient knowledge for real-world applications, which use extracted information to learn more from the previous experiences. As a result, the smart technologies, like IoT devices, have the opportunity to know or predict the real users' needs. However, we need to understand more the integration of Big Data with other promising technologies, such as IoT applications, smart cities, and so on.

Table 1: Clustering algorithms based on mapreduce.

Paper	Clustering Algorithm	Category
(Nguyen et al., 2013)	K-means	Partitional algorithm
(Lin et al., 2011)	Cop-k-means	
(Al-Madi et al., 2014)	PSO	
(Srirama et al., 2012)	PAM	Centroid-based clustering
(Ng and Han, 2002)	CLARA	
(Amini et al., 2014)	DBSCAN	Density-based algorithm
(Guo et al., 2015)	SCAN	
(Goyal et al., 2016)	GridSLINK	Single-linkage hierarchical clustering
(Rafailidis and Manolopoulos, 2017)	Spectral	
(Liu et al., 2014)	Co-clustering	Co-clustering
(Pandove and Goel, 2015)	CURE	Hierarchical algorithm
(Wang et al., 1997)	STING	Grid-based algorithm

3 RESEARCH CHALLENGES AND OPPORTUNITIES IN BIG DATA AND IoT

Based on the review of Big Data technologies and related clustering algorithms, we found that there is a set of research challenges that are worth to further investigate when linking Big Data clustering and IoT. It is hence valuable to develop a research agenda that contains possible research topics and questions to be addressed in the future. In order to connect IoT and clustering algorihtms in Big Data, we have initated the investigation from the direction of IoT data features. IoT data have certain specifics indicating that certain clustering algorithms can be expected to yield higher analytical effectiveness and efficiency. The key common characteristics of IoT data are as follows.

- *Homogeneity in heterogeneity*: Although there is an enormously large number of data sources in a typical IoT network (sensor devices) with

different types of IoT devices, there are many instances of each type (for instance many unified temperature sensors, smart utility meters).

- *Size of the data records:* The data records produced by IoT devices are typically very small and well structured.
- *Time series format:* The data records are typically produced in fixed time intervals and delivered in time series, which is a characteristic that can ease the analysis.
- *Low data quality given by device reliability:* There is often low pressure on the reliability of the devices, because the number of them is very high (it is beneficial to have many cheap sensors instead of a few expensive sensors), but with high pressure on elaborated techniques to combine the data from different devices that in a way substitute each other (e.g. missing information is deduced from neighbouring devices).
- *High security risks:* Each device might become a security vulnerability, hence all analytical methods should be robust against injected data by attackers (suspicious data should be continuously evaluated and not be considered in the analysis).
- *Dynamism:* It is a common characteristic of IoT networks that IoT devices dynamically join and leave the network.

The critical research question in this respect is how to optimize data clustering and analysis to take advantage of the characteristics of IoT data. Furthermore, IoT technologies have been incorporated into various important domains in our life. Thus, IoT domains refer to the IoT techniques that are applied in certain context such healthcare IoT or transportation IoT. Di□erent IoT domains share a set of common features but the the same time possess domain-specific variations. For example, most of the IoT domains emphasize the data collection, monitoring, sharing, automation, control and collaboration. Also, their datasets usually consist of relatively homogeneous data records e.g. from sensors and other IoT devices, which are often in a time series. However, healthcare or military IoT may acquire more precision or security and tranpotation or smart city IoT may have a relatively loose quality standard to the data.

3.1 Clustering Algorithms in Big IoT Data

Most of the IoT applications are based on the vision of connecting different objects to each other and analyze the generated IoT data in real-time. In this dynamic IoT context, the clustering algorithms are used to ensure the reliability of the IoT distributed applications. For example, in (Fredj et al., 2013), a hierarchical agglomerative clustering technique has been descripted to provisione a scalable search for constructing routing tables which perform request matching and forwarding.

The data from IoT devices are possibly generated from different sources such as sensors or mobile devices (Hossain et al., 2017). These data, termed as Big IoT Data in this paper, are with large volume, fast-moving and usually unstructured e.g. image data or stream data. The Big Data analytics mainly aims to firstly classify the Big Data, then mine the patterns and finally produce predictions. For example, in a city, there can be real-time traffic image data from various IoT devices such as road surveillance, satellite photos and traffic sensors etc. In order to analyze the real-time density of the cars, the traffic image data from different sources need to be firstly clustered and then processed for further analysis. However, for data clustering, the road surveillance images data and stream traffic sensor data are considered as one IoT data input but with different data structures. Maybe the combination of the clustering with other methods, like fuzzy, could improve further the functionality of clustering methods (D'Urso et al., 2017). However, the treatment of multi-structured data is still a big lack of clustering algorithms that based on an unsupervised process of classification of data into clusters.

Due to the large volume, complexity, variety, and rapid generation of IoT data that create an important opportunity in our daily life, new tremendous challenges have been raised for researchers to design new scalable and efficient clustering methods that are able to supervise or semi-supervise input data, classify multi-structured data, detect noisy data, and extract valuable information from different IoT data sources. We therefore tackle this critical problem by firstly reviewing the clustering algorithms in Big Data. Given the nature of the Big IoT Data, we have proposed the following research questions.

- How to effectively select the clustering algorithms for Big IoT Data?

- How to dynamically select the most suitable algorithm to cluster the Big IoT Data in a timely manner?
- How to cluster the different types of Big IoT Data that represent the same or similar entity/event?
- How to choose the proper Big Data technologies such as MapReduce frameworks to perform the clustering algorithm for Big IoT Data?

3.2 Dynamics of IoT Systems

There is high dynamics in IoT systems, especially for the mobile IoT system, where the mobile devices can be considered as a highly dynamic IoT end point. For example the IoT devices may spontanously join or leave the networks due to the mobility. In fact, mobile devices are sensors devices that are able to sense, process and generate large amount of real world data. Then the collected contextual information is analyzed, interpreted and utilized to make decision in different areas i.e. business intelligence [46]. Nowadays, most persons are surrounded by multiple mobile devices that can provide several services to the end-user at any time and place. Due to the increasing impact of mobile devices on people's habitudes, mobile devices have therefore become a major part of the IoT paradigm (Ahmed and Rehmani, 2017).

Furthermore, the characteristics of mobile devices, notably mobility, enhance further the integration of mobile devices in IoT by offering a wide range of promising innovations, which will dramatically make the data more valuable for several domains in the forthcoming years, such as education, healthcare, smart homes, smart cities, and so on. However, the mobile Big Data require dynamic analysis techniques for ensuring the usability of data. Yet, the clustering algorithms could be thought of as a key to solve this challenge, such as the application of K-Medoids clustering in (Dash et al., 2012), where the algorithm has been used as a way of facilitating privacy for organized data in mobile cloud computing.

Our review found that there is an essential need for new techniques related to different aspects of testing and quality evaluation of mobile Big Data. Accordingly, the success of mobile Big Data leads to these relevant research questions.

- How to use context-aware, location-aware and users' experience to test and evaluate the compatibility and adaptability of data?
- How could mobile structured and unstructured data be used effectively?
- What are the issues facing organizations trying to take the benefits of mobile Big Data?
- What are the limitations of the existing analytical methods, especially clustering models, to process mobile Big Data? And what are the new methods in response to these issues?
- What are the best practices and strategies that the organization need to adopt for Big Data projects?

3.3 The Role of Networking in IoT

The increasing number of IoT devices, including mobile devices, has increased the amount of data exponentially. As a result, the 5th generation mobile network (5G), which is expected to be operational by 2020, has become a key success factor to support various types of emerging IoT applications with strengthened quality of service (Jiang and Liu, 2017). To achieve more efficient communication between IoT sensors in 5G. The clustering techniques have been used (Xu et al., 2017). However, the 5G must be more smart and flexible to guarantee the quality of their services for the end users as well as for the smart environments. In other words, the 5G has to learn from known and unknown data in order to achieve their goals, which are share data everywhere, every time, by everyone and everything for the benefits of several domains such as healthcare, business, and so on. Yet, the advancement of 5G networks conducts to these relevant research questions.

- How could the 5G and other modern networks handle the real-time and online mobile-data processing?
- What are the most important factors that need to be taken into consideration when designing and evaluating solutions for Big Data and IoT technologies in 5G and other modern mobile communication systems?
- What is the impact of machine learning in 5G mobile information systems? And how could data mining supply the advancement of 5G mobile information systems?

3.4 Machine Learning Applications to Support IoT

Since machine learning techniques, such as supervised and unsupervised learning, tend to classify and cluster the data, thus, different machine learning

models may bring insightful results for IoT data analytics. In general, machine learning is an interdisciplinary approach in building mathematical methodologies that are ideally suited to the task of extracting knowledge from Big Data and make data-driven predictions and decisions. It can be defined as: "a field of study that gives computers the ability to learn without being explicitly programmed" (El Naqa and Murphy 2015). In other words, the goal of machine learning is to develop algorithms that can be self-programmed to solve new problems by using previous known data rather than directly programming new solutions. In the IoT context, the machine learning techniques have been linked to clustering algorithms for deep learning because the clustering concept is an unsupervised process of classification of data into groups. For example, in (Akbar et al., 2015), the machine learning methods have been used to analyze automatically the IoT data based on real-time rules. Meanwhile, the integration between Big Data and mobile Internet can produce positive impacts in the machine learning field by providing more real world examples to forecast the future activities and investments.

We found that it is valueable to develop new technologies, methods and algorithms for this explosive increase of big masses of mobile IoT data that need more real-time analysis which challenges the existing traditional analytic tools as well as the existing machine learning algorithms. Therefore, we identified the following research questions associated with the machine learning applications with support of mobility patterns.

- How could machine learning applications improve the existing clustering models for self-mobile-IoT applications?
- How could machine learning optimize mobile Big Data as a service and analytics as a service for the IoT?
- What are the best strategies to process the mobile Big Data and extract the most useful information?

4 CONCLUSIONS

In this paper, we have conducted a survey on Big Data technologies and clustering algorithms, in which we have specified the pros and cons of each algorithm in the Big Data context. We have further related our review to the research of IoT and discussed the relations between Big Data, clustering algorithms and IoT. Based on the review, we have proposed a set of research challenges that address the emerging research topics and research questions on the data clustering in IoT, dynamics of Big Data application in IoT, the role of networking in IoT, as well as the machine learning applications.

The research challenges can be considered as a research agenda to guide the future research across Big Data and IoT communities. Specifically, this paper has emphasized the importance of clustering algorithms in Big IoT Data and brings attention and possible applications of the Big Data clustering algorithms for IoT. As future work, we plan to further investigate the relations between the communities of Big Data and IoT. We will analyze which Big Data technologies can be effectively used in which IoT context. Also, we plan to investigate data features from IoT and connect them with features in Big Data, which can facilitate Big Data applications in IoT.

ACKNOWLEDGEMENTS

The work was supported from European Regional Development Fund Project "CERIT Scientific Cloud" (No. CZ.02.1.01/0.0/0.0/16_013/0001802).

REFERENCES

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

Amini, Amineh, Teh Ying Wah, and Hadi Saboohi. "On density-based data streams clustering algorithms: a survey." Journal of Computer Science and Technology 29.1 (2014): 116-141.

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

Ahmed, Ejaz, and Mubashir Husain Rehmani. "Mobile edge computing: opportunities, solutions, and challenges." (2017): 59-63.

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

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, no. 1 (2016): 1-42.

Da Xu, Li, Wu He, and Shancang Li. "Internet of things in industries: A survey." IEEE Transactions on industrial informatics 10.4 (2014): 2233-2243.

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 16(1), 2017, 51-70.

Dash, Sanjit Kumar, Debi Pr 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.

Dey, Anind K. "Understanding and using context." Personal and ubiquitous computing 5.1 (2001): 4-7.

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

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 2(3) (2014): 267-279.

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.

Guo, Kun, Wenzhong Guo, Yuzhong Chen, Qirong Qiu, and Qishan Zhang. "Community discovery by propagating local and global information based on the MapReduce model." Information Sciences 323 (2015): 73-93.

Goyal, Poonam, 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.

Havens, Timothy C., 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.

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 55(1) , 2017, 18-19.

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

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

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 15, no. 1 (2014): 37.

Lin, Chao, Yan Yang, and Tonny Rutayisire. "A parallel Cop-Kmeans clustering algorithm based on MapReduce framework." Knowledge Engineering and Management. 2011. 93-102.

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." Physica A: Statistical Mechanics and its Applications (2017).

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.

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

Mohebi, Amin, Saeed Aghabozorgi, Teh Ying Wah, Tutut Herawan, and Ramin Yahyapour. "Iterative Big Data clustering algorithms: a review." Software: Practice and Experience 46, no. 1 (2016): 107-129.

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.

Ng, Raymond T., and Jiawei Han. "CLARANS: A method for clustering objects for spatial data mining." IEEE transactions on knowledge and data engineering 14.5 (2002): 1003-1016.

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.

Rafailidis, D., E. Constantinou, and Y. Manolopoulos. "Landmark selection for spectral clustering based on Weighted PageRank." Future Generation Computer Systems 68 (2017): 465-472.

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.

Srirama, Satish Narayana, Pelle Jakovits, and Eero Vainikko. "Adapting scientific computing problems to clouds using MapReduce." Future Generation Computer Systems 28.1 (2012): 184-192.

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

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

Wang, Wei, Jiong Yang, and Richard Muntz. "STING: A statistical information grid approach to spatial data mining." VLDB. Vol. 97. 1997.

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 4, no. 5 (2017): 1229-1249.