Iot Data Preprocessing - A Survey

V.A. Jane[1] , Dr. L. Arockiam[2]

[1, 2]Department of computer Science, St. Joseph's College (Affiliated to Bharathidasan University), Trichy, Tamilnadu, India 62001.

Abstract

The Internet of Things (IoT) is a rapidly evolving system in engineering and science. The sensors utilized in numerous sectors output a significant quantity of data. As a result, several consumers have a clear desire for efficient knowledge from these vast databases. This enormous dataset is far from flawless; it has several flaws (like distortion, inconsistent data, and anomalies) and is unsuitable for investigation due to the risk of inaccurate results. As a result, data preprocessing is a necessary approach for these information. Data preprocessing is a crucial and necessary phase, with the primary purpose of using procedures to filter, refine, repair, and enhance the raw information. The purpose of this study is to conduct a survey of IoT data preprocessing and methodologies. This article analyses current data preprocessing studies in the IoT environment, as well as the history of IoT data preprocessing and review articles of sophisticated data preprocessing approaches. The image clearly depicts the categorization of different preprocessing methodologies and procedures. Preprocessing cleaning, conversion, minimization, and integration methods are discussed. Furthermore, strategies for implementing such ideas in IoT data preprocessing are presented. IoT approaches for data preparation in diverse applications are listed. Lastly, difficulties and obstacles are explored that will be valuable in future research.

Key Words IoT, Preprocessing, Data Cleaning, Noise handling

Introduction

The Internet of Things (IoT) is a system of items that are linked to the Internet. It's a powerful automated and intelligence platform with applications in a variety of sectors and distinctive adaptability and capabilities in every provided setting (for example, agriculture and medicine) [1]. Being linked to the Internet, one may gather information and transfer it over the web, acquire data from the web, or do both. In the IoT, sensors and other devices produce data tremendously. Such information are transported to the cloud for processing, evaluation or modeling and to construct software applications. Big data analysis is a very essential method of finding insights from such information [2]. Data preparation is essential before evaluating the data since it contains various flaws like missing data, distortion, and inconsistency. One of the most important aspects of the knowledge discovery process is data preparation [3]. Low-quality information may weaken the efficacy of subsequent learning algorithms. As a result, limiting the effect on reliability improves the dependence of following automated findings and improves judgments via the use of appropriate processing techniques. Data transformation, data reduction, data standardization, data cleansing, and data integration are some of the strategies used [5]. By separating complicated continual feature sets and choosing and deleting undesired and noisy characteristics, such strategies minimize the information. Throughout this procedure, the actual construction of the data must be preserved while a more acceptable size is achieved. Quick training of learning approaches, sophisticated generalization abilities, and improved comprehension and convenient analysis of the results are among the advantages of data processing [6]. The purpose of this study is to conduct a survey of data preprocessing, its approaches, and current data preprocessing achievements. The following is the framework of this work: Firstly, data pretreatment ideas in IoT contexts (part 2), as well as data preprocessing methodologies. Section 3 explains the methods used in numerous IoT-based applications, and Section 5 brings this project to a close.

Related work

Physical sensor faults that arise throughout the data gathering process were examined by Hui et al., [7]. This article describes wide range of physical sensor discrepancies, error - detecting processes, and error - correcting methods, as well as the variations between them. Principal component analysis (PCA) and Artificial Neural Network (ANN) were the best error-detection and rectification procedures.

Mathew et al., [8] examined Kalman filter, z-scoring, and moving Average filter processing approaches. To begin, the chemical sensor data was cleaned using pre-processing steps. Following that, the information is cleaned and assessed utilizing classification techniques like Linear Discriminant Analysis (LDA), K Nearest Neighbor (KNN), and Support Vector Classifier (SVC). Lastly, the different preprocessing approaches' efficiencies were evaluated. Among them, the Kalman filter approach was shown to produce superior results than the rest.

For a higher-dimensional Microarray tumor sample, Zena et al. [9] analyzed feature-selection and feature-extraction approaches. In the micro array set of data, the researcher addressed the implications of redundant and irrelevant attributes. The significance of dimension reduction, as well as its benefits and downsides, also were explored.

The method of data transfer in an IoT context was outlined by Chao et al.,[10]. Although a pretreatment approach was employed to reduce transmission time and increase processing speed, this research centered on the latter.

Evgeniy [11] suggested a processing system for sensor information. Several processing strategies that are appropriate for the proposed framework have been discovered. This framework used streaming sensor data from the Univariate time series dataset.

Filter-based monitoring solution for IoT environment was suggested by Natarajasivan et al., [12]. Accelerometers, position sensors, vision sensors, audio sensors, temperature sensors, and directional sensor readings were all used in this project. The obtained information from such sensing devices was processed utilizing Kalman filter, and the performance was evaluated employing SVM. The suggested scheme took longer to complete.

Cleber et al., [13] conducted a review of all IoT application journal articles since 2015. The authors assigned a value to the IoT application depending on how it was used. When contrasted to other apps, smart home applications are commonly employed by investigators. In addition, the sensor used in intelligent devices is explored.

Rajalakshmi et al., [14], addressed the concept of IoT in intelligent systems and summarized sensor data-collection issues like data aggregation, extensibility, data fusion, de-noising, variability, data anomaly analysis, real-time computation, and missing data imputation. The authors discuss the IoT data analytics procedure using a drone for a traffic-monitoring scheme and described how cloud, fog, and edge computing are used in IoT to enhance the analytics platform.

Data gathering, cleansing, data aggregation, data migration to the cloud, and data processing were all discussed by David et al., [15] in their examination of data management issues in the IoT context. AI, machine learning, deep learning, and data mining are some of the enhanced data-processing innovations explored by the researcher.

Karinaer al., [16] gave an overview on processing strategies as well as data mining-related challenges. The essential ideas of data mining, as well as processing approaches and challenges, were thoroughly addressed. It also explored recommendations for future and provided alternative ideas.

Data preprocessing methodologies for the big data era were suggested by Garca et al., [17]. The critical elements of data processing were discussed, as well as the existing key problems. In addition, various data preprocessing techniques for text mining were examined, including discretization and normalization, extraction of features, feature selection, feature indexers and encoders, and other methods. The importance of large data preparation was also stressed.

In the area of data mining, Jayaram et al., [18] published a survey on data preparation strategies. The major goal was to find answers to different data preparation issues. The authors concentrated on data cleaning techniques such as filtering, imputation, hybrid approach, wrapper techniques, and embedded methodologies. Every technique's procedure and applications were detailed with instances. Distortion and data management, particularly, were discussed, as well as instructions on how to identify and handle it. Lastly, the difficulties encountered while performing data cleansing in various domains were shown.

Several large data processing strategies were explored by Huma Jamshed et al., [19] to cleanse input for subsequent mining and analytical jobs. The basic phases of data preparation, such as data filtering, data conversion, data reduction, and data aggregation, were first described. Following that, architecture for internet data preparation was presented, with every step described in detail. Lastly, the model was applied to the basic textual data, and processing stages such as removal of noise, tokenization, and normalization were completed.

Taxonomy of preprocessing techniques

Data processing is the process of preparing actual data to be used in data mining [20]. Actual data is noisier, has incomplete data, it includes a lot of uncertain information, and it's enormous. Such factors influence the quality of the data to deteriorate during the mining or modeling process, resulting in poor results. As a result, information should be improved before it can be mined or modeled. This is referred to as data preprocessing. There are several approaches for doing such operations in order to make the data appropriate for analysis. IoT data preprocessing techniques are shown in figure1.

Data Cleaning

Data cleaning [21] can be defined as the process of eliminating the erroneous and missing part in the data. The process of handling these noisy and missing values can be achieved by various ways.

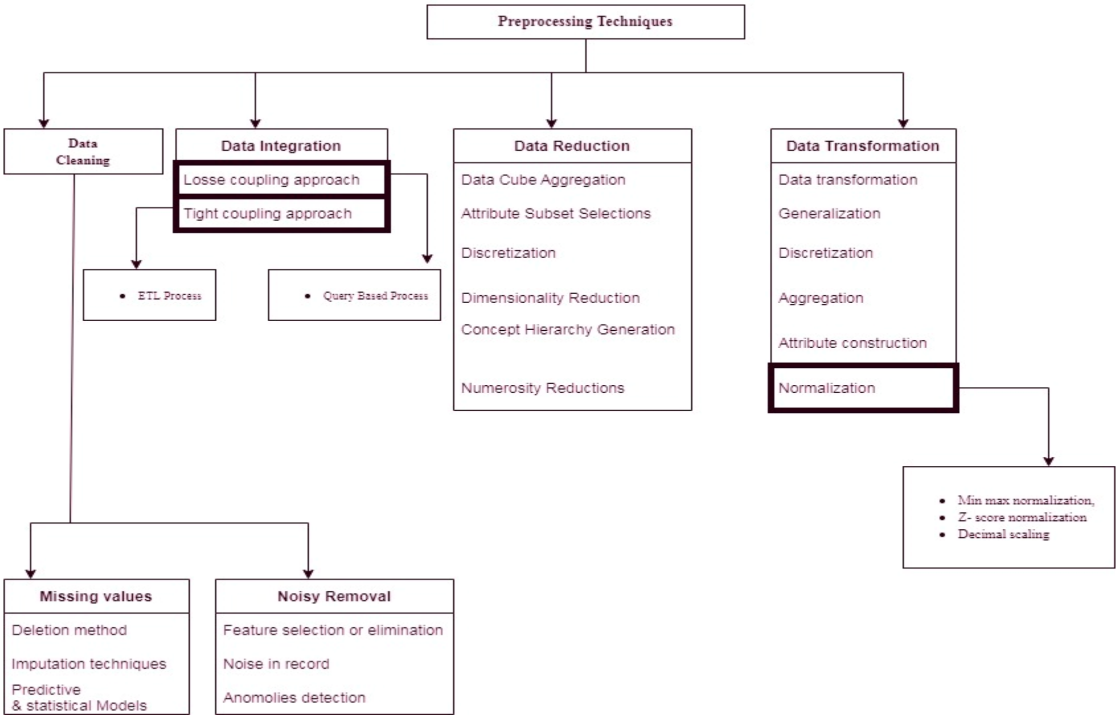

Fig 1: Taxonomy of IoT data preprocessing

Deal with missing values

Now a days, Missing values are present in all types of datasets that from existing, industrial and devices [22]. They have different reasons like manual data entry procedures, equipment Errors and misalignments. Such datasets require a pre-processing level to prepare and clean a data for effective and sufficient knowledge extraction process. There are machine learning algorithms and packages that can automatically detect and manipulate missing data, but it is recommended to manually replace missing data with analysis and coding techniques [23]. First, the types of missing have to understand and basically missing is classified into three types that are missing at random (MCAR), missing at Random (MAR) and missing not at Random (MNAR) [24]. To deal all these types of missing values using various methods like removing data [25], imputation techniques [26], multiple imputation techniques [27], statistical or predictive models [28]. In the removing method, missing values handle through list wise, pairwise and dropping variables techniques. The goal of any imputation technique is to create a complete dataset, which can be used for machine learning. Some of the imputation techniques are, deductive imputation, mean/ median /mode imputation, random sampling imputation, regression imputation, multiple imputation [29]. Predictive and statistics model also used to impute the missing data. Most commonly used in this process are, linear regression, Random forest, k nearest neighbor, expectation maximization and sensitivity analysis [30].

Deletion / Removing Method

Deletion / removing methods used to remove the missing values using various approaches that are following,

List-wise deletion

The most common approach to handle missing data, it simply omit the data that with missing values after analyze the remaining data. This process also called as complete case analysis [31]. If the samples are large and the assumption of MCAR is satisfied, list wise deletion is in reasonable strategy. However, when a large sample is not available, or the assumption of MCAR is not satisfied, the list wise approach is not the optimal strategy. This approach establishes the bias if it does is not fulfill the MCAR.

Pairwise Deletion

Pairwise deletion technique remove missing observations only and the existing variables are analyzed [32]. If there is no data elsewhere in the data set, existing values are used. This approach uses all of the observed information, so it saves more information than list wise deletion technique. Pairwise deletion also called as Available Case Analysis (ACA). Pairwise deletion is known to be less dependent on MCAR or MAR data. However, if there are many missing observations, the analysis is flawed. The problem with pairwise elimination is that even if one takes the available cases, one cannot compare the analyzes because each time the model is different.

Dropping Variables

Dropping variables approach removes a variable or column from a dataset if it contains more missing values [33]. This approach is performs depend on the situation there is no rule to handle this approach and requires a proper analysis of the data before the variable is dropped all together. This should be the last option to test whether the model improves performance after the variable is removed.

Dealing with Noisy data

Noisy data is an unwanted data item, feature, or record that does not help explain the feature, or the relationship between the feature and the target [34]. Noisy data can affect the algorithms to find the patterns in the data. Noise can be classified into three types [35]. Noise 1 is anomalies in some data items, noise 2 is features that don’t help to the target like irrelevant or weak features, and noise 3 is which records that do not follow the form or relationship that like the rest of the records. If the noise is in the features, feature selection or elimination techniques to best for handling noise in the features this includes filter method, wrapped method and embedded methods. For handling noise in records, k fold validation and manual methods are used in basic. Unsupervised methods are used to detect anomalies in data items. Some of them are, density based anomaly detection, cluster based anomaly detection and SVM based anomaly detection [36].

Data integration

Data integration **[37-38]** is one important techniques in preprocessing which combines data from different source and giving users an integrated view of this data. Data sources may contain databases, data cubes, or flat files. One of the most popular implementations of data integration is the creation of a company's data warehouse. Mainly, Data integration is done through two main approaches known as the "tight coupling approach" and "loose coupling approach"[39]. Tight coupling defines Data from various sources that combined into one place by the process of Extraction, Transformation and Loading. Single physical location provides a balanced interface for querying data and ETL process provide identical data warehouse. Loose coupling data exists only in real source databases. In loose coupling, virtual mediation schema takes an interface from the user to the query, converts and sends it to a source database for getting result. As well as there are many "adapters" or "wrappers" in the mediation schema that can be reconnected to the source systems and bring the data to the front end.

Data reduction

Over the past decades, data generation and storage in data bases or data warehouses has increased. So, these amounts data can take a very long time to perform data analysis and mining process. Data reduction **[**40-41**]** techniques can be used to obtain a data set, which are very small in size but yield the same analytical results. Traditional, data reduction approaches [42-43] are Data cube aggregation, Attribute subset selections, Dimensionality reduction, Discretization and concept hierarchy generation and Numerosity reductions. Data cube aggregation used to construct a data in simple form. It applied on the data and form a data cubes. Attribute subset selection technique remove irrelevant, weakly features or redundant features. This process can be achieved by various statistical and computational methods like filter methods, wrapper methods, and embedded methods. Dimensionality reduction is the reduction technique to reduce the size of dataset. This process used to reduce the number of random variables to be considered by obtaining the set of the principal variable. Dimensionality reduction reduces the amount of data by eliminating outdated or unwanted features using techniques that includes PCA, backward feature elimination, forward feature construction, and discriminant methods. Since real data is replaced with real data, with mathematical models or a small representation of data like parameters or non-parametric method such as clustering, sampling and histogram. Discretization & Concept Hierarchy Operation techniques are used to change the raw data values for the attributes by a range or by more conceptual conditions. This is a form of numerical reduction, which is very useful for the automatic generation of concept sequences. Discretization techniques follows two ways namely top down discretization and bottom up discretization. Concept hierarchies for numeric data that includes techniques are binning, histogram analysis and clustering.

Data transformation

Data transformation [44] is the process of converts' data from one format to another format. Data transformation includes smoothing, aggregation, discretization, attribute construction, normalization and generalization. Smoothing used to remove noise from a dataset though various algorithms and highlight the significant features in a dataset. Data normalization involves converting all data variables into a specific range. Techniques used for normalization min max normalization, z- score normalization and decimal scaling**.**

Conclusion

Big data is currently widely used in a variety of fields, including academia, agribusiness, medicine, organizations, and web mining. Studying from such vast amounts of data is both an exciting and difficult undertaking. Information acquired from massive quantities of data offers tremendous prospects and has the ability to alter several industries. However, since big data contains imperfections such as noise and incomplete information, it may reduce the effectiveness and reliability of decision-making. As a result, data refining is required. In the case of larger data settings, this work presents a systematic flow of research on data preparation strategies. The principles of data preparation were discussed, as well as literature evaluations pertaining to data preprocessing approaches. The image clearly demonstrated the categorization of different pre-processing methodologies and procedures. Preprocessing, cleansing, transformations, reductions, and collaboration with methodologies and procedures were all shown. On a variety of applications, data preparation methods were tabulated. Finally, difficulties and concerns that should be addressed in the future were discussed.

References

[1] Bramer, Max. “Data for data mining”, In Principles of data mining”, pp. 9-19. Springer, London, 2016.

[2] Alasadi, Suad A., and Wesam S. Bhaya.”Review of data preprocessing techniques indata mining, Journal of Engineering and Applied Sciences 12, no. 16 (2017): 4102-4107, 2017.

[3] Cordón, Ignacio, Julián Luengo, Salvador García, Francisco Herrera, and Francisco Charte. & quot; Smart data: Data preprocessing to achieve smart data in r.& quot; Neuro computing 360, 1-13, 2019.

[4] Hu, Hanqing, and Mehmed Kantardzic. & quot; Smart preprocessing improves data streammining. & quot; In 2016 49th Hawaii International Conference on System Sciences (HICSS), pp.1749-1757. IEEE, 2016.

[5] Shi, F.; Li, Q.; Zhu, T.; Ning, H., “A survey of data semantization in internet of things”, Sensors, 18, 313, 2018.

[6][24] Shah, S. H., &Yaqoob, I, “A survey: Internet of Things (IOT) technologies, applications and challenges“, IEEE Smart Energy Grid Engineering SEGE). doi:10.1109/sege.2016.7589556, 2016.

[7] Teh, Hui Yie, Kempa-Liehr, Andreas W, Wang, Kevin I-Kai, "Sensor data quality: a systematic review", Journal of Big Data, 7(1), 11–60, 2020, doi: 10.1186/s40537-020-0285-1

[8] Weiss, Matthew, Wiederoder, Michael S, Paffenroth, Randy C, Nallon, Eric C, Bright, Collin J, Schnee, Vincent P, McGraw, Shannon; Polcha, Michael, Uzarski, Joshua R, "Applications of the Kalman Filter to Chemical Sensors for Downstream Machine Learning", IEEE Sensors Journal, (), 1–1, 2018, doi:10.1109/JSEN.2018.2836183

[9] Hira, Z. M., &Gillies, D. F, "A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data", Advances in Bioinformatics, 1–13, 2015,doi:10.1155/2015/198363

[10] Xu, C., Yang, H. H., Wang, X., &Quek, T. Q. S, "On Peak Age of Information in Data Preprocessing enabled IoT Networks", IEEE Wireless Communications and Networking Conference (WCNC), 2019, doi:10.1109/wcnc.2019.8885690

[11] Evgeniy Latyshev, "Sensor Data Preprocessing, Feature Engineering and Equipment Remaining Lifetime Forecasting for Predictive Maintenance", Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL'2018), 226-231, 2018.

[12] D. Natarajasivan and M. Govindarajan, "Filter Based Sensor Fusion for Activity Recognition using Smartphone", International Journal of Computer Science and Telecommunications Volume 7, Issue 5, 2016.

[13] Morais, C. M. de, Sadok, D., & Kelner, J, "An IoT sensor and scenario survey for data researchers", Journal of the Brazilian Computer Society, 25(1), doi:10.1186/s13173-019-0085-7, 2019.

[14] Rajalakshmi Krishnamurthi, Adarsh Kumar, Dhanalekshmi Gopinathan, Anand Nayyar, and Basit Qureshi, "An Overview of IoT Sensor Data Processing, Fusion, and Analysis Techniques", Sensors, 20, 6076; doi:10.3390/s20216076.

[15] Gil, D., Johnsson, M., Mora, H., & Szymanski, J, "Advances in Architectures, Big Data, and Machine Learning Techniques for Complex Internet of Things Systems", Complexity, 1–3, doi:10.1155/2019/4184708,2019.

[16] Gibert, Karina, Miquel Sànchez–Marrè, and Joaquín Izquierdo. "A survey on pre-processing techniques: Relevant issues in the context of environmental data mining." AI Communications 29, no. 6 (2016): 627-663.

[17] García, Salvador, Sergio Ramírez-Gallego, Julián Luengo, José Manuel Benítez, and Francisco Herrera. "Big data preprocessing: methods and prospects." Big Data Analytics 1, no. 1 (2016): 9.

[18] Hariharakrishnan, Jayaram, S. Mohanavalli, and KB Sundhara Kumar. "Survey of pre-processing techniques for mining big data." In 2017 International Conference on Computer, Communication and Signal Processing (ICCCSP), pp. 1-5. IEEE, 2017.

[19] Jamshed, Huma& Khan, M. &Khurram, Muhammad & Inayatullah, Syed &Athar, Sameen. (2019). Data Preprocessing: A preliminary step for web data mining. 206-221. 10.17993/3ctecno.2019.specialissue2.206-221.

[20] [21] Shobanadevi, A., & Maragatham, G. Data mining techniques for IoT and big data — A survey, "International Conference on Intelligent Sustainable Systems (ICISS)", ISBN:978-1-5386-1959-9,doi:10.1109/iss1.2017.8389260,2017.

[21] Syaifudin, Yan Watequlis, and Dwi Puspitasari. "Twitter data mining for sentiment analysis on people's feedback against government public policy." MATTER: International Journal of Science and Technology 3, no. 1, 2017.

[22] Fatima, Meherwar, and Maruf Pasha. "Survey of machine learning algorithms for disease diagnostic" Journal of Intelligent Learning Systems and Applications, 9, no. 01, 1, 2017.

[23] Yadav, Madan Lal, and Basav Roychoudhury. "Handling missing values: A study of popular imputation packages in R", Knowledge-Based Systems, 160, 104-118, 2018.

[24] Gomes, Harold, "Evaluation of Patterns of Missing Prices in CPI Data.", 2018.

[25] Huang, Min-Wei, Wei-Chao Lin, Chih-Wen Chen, Shih-Wen Ke, Chih-Fong Tsai, and William Eberle. "Data preprocessing issues for incomplete medical datasets" Expert Systems, 33, no. 5, 432-438, 2016.

[26] Wang, Guang C., Kenny C. Gross, and Dieter Gawlick. "Missing value imputation technique to facilitate prognostic analysis of time-series sensor data." U.S. Patent Application 16/005,495, filed December 12, 2019.

[27] Chhabra, Geeta, Vasudha Vashisht, and Jayanthi Ranjan, "A comparison of multiple imputation methods for data with missing values", Indian Journal of Science and Technology, 10, no. 19 (2017): 1-7.

[28] Alexandropoulos, Stamatios-Aggelos N., Sotiris B. Kotsiantis, and Michael N. Vrahatis. "Data preprocessing in predictive data mining." The Knowledge Engineering Review, 34, 2019.

[29] Lin, Wei-Chao, and Chih-Fong Tsai. "Missing value imputation: a review and analysis of the literature (2006–2017)" Artificial Intelligence Review, 53, no. 2 (2020): 1487-1509, 2020.

[30] Gad, Ibrahim, and B. R. Manjunatha. "Performance evaluation of predictive models for missing data imputation in weather data." In 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1327-1334. IEEE, 2017.

[31] Hariharakrishnan, Jayaram, S. Mohanavalli, and KB Sundhara Kumar. "Survey of pre-processing techniques for mining big data." In 2017 International Conference on Computer, Communication and Signal Processing (ICCCSP), pp. 1-5. IEEE, 2017.

[32] Al-Utaibi, Khaled A., and El-Sayed M. El-Alfy. "Intrusion detection taxonomy and data preprocessing mechanisms", Journal of Intelligent & Fuzzy Systems 34, no. 3 (2018): 1369-1383, 2018.

[33] Zulkepli, Fatin Shahirah, Roliana Ibrahim, and Faisal Saeed. "Data preprocessing techniques for research performance analysis." In Recent Developments in Intelligent Computing, Communication and Devices, pp. 157-162. Springer, Singapore, 2017.

[34] Misra, Puneet, and Arun Singh Yadav, "Impact of Preprocessing Methods on Healthcare Predictions" Available at SSRN 3349586 (2019).

[35] Nayak, Arjun Srinivas, A. P. Kanive, Naveen Chandavekar, and R. Balasubramani. "Survey on pre-processing techniques for text mining." International Journal Of Engineering And Computer Science, ISSN (2016): 2319-7242, 2016.

[36] Kumar, HM Keerthi, and B. S. Harish. "Classification of short text using various preprocessing techniques: An empirical evaluation" In Recent Findings in Intelligent Computing Techniques, pp. 19-30. Springer, Singapore, 2018.

[37] Hui, Jingya, Lingli Li, and Zhaogong Zhang. "Integration of big data: a survey." In International Conference of Pioneering Computer Scientists, Engineers and Educators, pp. 101-121. Springer, Singapore, 2018.

[38] Samoilova, Evgenia, Florian Keusch, and Tobias Wolbring. "Learning analytics and survey data integration in workload research." Zeitschrift für Hochschulent wicklung. Special Edition: Learning Analytics: Implications for Higher Education 12, no. 1 (2017): 65-78.

[39] Osial, Phillip, Kalle Kauranen, and Emdad Ahmed. "Smartphone recommendation system using web data integration techniques." In 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), pp. 1-5. IEEE, 2017.

[40] ur Rehman, Muhammad Habib, Chee Sun Liew, Assad Abbas, Prem Prakash Jayaraman, Teh Ying Wah, and Samee U. Khan. "Big data reduction methods: a survey." Data Science and Engineering 1, no. 4 (2016): 265-284, 2016.

[41] ur Rehman, Muhammad Habib, Victor Chang, Aisha Batool, and Teh Ying Wah. "Big data reduction framework for value creation in sustainable enterprises." International Journal of Information Management 36, no. 6 (2016): 917-928, 2016.

[42] Weng, Jiaying, and Derek S. Young, "Some dimension reduction strategies for the analysis of survey data", Journal of Big Data, 4, no. 1 (2017): 1-19, 2017.

[43] Jindal, Priyanka, and Dharmender Kumar. "A review on dimensionality reduction techniques", International journal of computer applications 173, no. 2 (2017): 42-46, 2017.

[44] Jarmin, Ron S., and Amy B. O'Hara. "Big data and the transformation of public policy analysis", Journal of Policy Analysis and Management, 35, no. 3 (2016): 715-721, 2016.