This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.

Jaribion, Alireza; Haghighat Khajavi, Siavash; Hossein Motlagh, Naser; Holmström, Jan
A Novel Method for Big Data Analytics and Summarization based on Fuzzy Similarity Measure

Published in:
Proceedings - IEEE 11th International Conference on Service-Oriented Computing and Applications, SOCA 2018

DOI:
[10.1109/SOCA.2018.00039](https://doi.org/10.1109/SOCA.2018.00039)

Published: 01/01/2018

Document Version
Peer reviewed version

Please cite the original version:
Jaribion, A., Haghighat Khajavi, S., Hossein Motlagh, N., & Holmström, J. (2018). A Novel Method for Big Data Analytics and Summarization based on Fuzzy Similarity Measure. In *Proceedings - IEEE 11th International Conference on Service-Oriented Computing and Applications, SOCA 2018* (pp. 221-226). [8599601] IEEE. <https://doi.org/10.1109/SOCA.2018.00039>

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.

A Novel Method for Big Data Analytics and Summarization based on Fuzzy Similarity Measure

Alireza Jaribion, Siavash H. Khajavi, Naser Hossein Motlagh, and Jan Holmström

Department of Industrial Engineering and Management

School of Science, Aalto University, Espoo, Finland

Emails: firstname.lastname@aalto.fi

Abstract—The use of Internet of Things (IoT) sensors and devices is on a sharp rise with the help of infrastructures like 5G cellular networks. However, with this swift expansion come major challenges such as management of the vast data collected by these IoT devices. In the paper, we propose a method which takes advantage of fuzzy similarity to significantly simplify the big data analysis and management for human operators and machines. A use case in the area of smart buildings was utilized to illustrate its potential application. A comparison is made between the ideal situation and the data collected from an office room. The considered environmental factors in this research are temperature, humidity, and workplace lighting and then the collected data was utilized to make triangular fuzzy numbers and after that, we compare them with an efficient fuzzy similarity measure. In addition to data summarization and abstraction, this method also protects the main information from inaccuracy. The advantage of the fuzzy controlling system is its aptitude to deal with nonlinearities and uncertainties.

Index Terms—Internet of Things, big data summarization, fuzzy similarity measure, environmental factors.

I. INTRODUCTION

The growing Internet of Things movement continues and it is expected to be 20 billion internet-connected things by 2020 [1]. With the improvements in internet technologies and wireless sensor networks (WSN), IoT is expanding as a ubiquitous global computing network where everyone and everything will be connected to the internet [2]. The number of devices using internet services is rising every day and having all of them connected by wire or wireless will put a comprehensive source of raw data at our fingertips. The concept of enabling interaction between machines is a cutting-edge technology but the technologies composing the IoT are not something novel [3]. According to Farooq et al. “IoT is the approach of converging data obtained from different kinds of things to any virtual platform on existing Internet infrastructure” [3]. Another definition for the Internet of Things according to Madakam et al. is “An open and comprehensive network of intelligent objects that have the capacity to auto-organize, share information, data, and resources, reacting and acting in face of situations and changes in the environment” [4].

In the light of an explosive data growth, the term “Big data” represents “the information assets characterized by such a high volume, velocity, and variety to require specific technology and analytical methods for its transformation into value” [5]. IoT is an important source that produces big data. Among smart cities constructed based on IoT, big data may come from

industry, agriculture, traffic, transportation, medical care, public and private facilities, etc. According to the processes of data acquisition and transmission in IoT, its network architecture may be divided into three layers: the sensing layer, the network layer and the application layer [6]. The sensing layer is responsible for data acquisition and mainly consists of sensor networks. The network layer is responsible for information transmission and processing, where close transmission may depend on sensor networks, and remote transmission shall depend on the Internet. Finally, the application layer supports specific applications of IoT. The data generated from IoT devices have a number of characteristics such as effective data accounts for only a small portion of the big data. It means a great number of noises may occur during the acquisition and transmission of data in IoT [6].

With data being a main focal point and core of IoT, algorithms and solutions must find ways to deal with big data from several sources with varying levels of integrity. Large amounts of data need to be processed by means of more complex methods than the usual statistical procedures [5]. One of the biggest challenges facing the IoT is data summarization for easy access to significant data. Fortunately, new summarization tools are making it easier, so the propagation of the IoT will be the next stage of the big data revolution [7]. During the current rapid development of IoT, big data, and cloud computing, the challenge is to explore big data and excerpt useful information or knowledge for action. Some applications require data to be accessed, extracted and processed in real-time. This can be challenging while dealing with big data to analyze data sets that are terabytes or petabytes in size.

Data summarization has become one of the indispensable tools and links in IoT similar to data mining. It has brought about a number of technical challenges and issues. According to Zhang et al. [8], the following matters for data mining can be sort into four categories such as real-time high-performance data mining, distributed data mining, data quality control, and decision-making control. Data summarization should facilitate data mining in these categories.

The fuzzy set theory has been used more and more frequently in intelligent systems because of its simplicity and similarity to human reasoning [9]. Several fuzzy learning algorithms for inducing rules from given sets of data have been designed and used to good effect with specific domains. Fuzzy logic is appropriate for the monitoring and fault detection

because quantitative concepts are often used for anomaly detection, and this is while fault detection itself involves fuzziness [10].

In this paper, a method was introduced which aims to streamline the transition from raw data collected from IoT devices towards the information usable for decision making. In other words, this method summarizes the data and help the decision maker to act appropriately.

In order to detect anomalous behavior, we utilize the data to make generalized fuzzy numbers and compute the similarity between the current status and ideal status with an appropriate fuzzy similarity measure. If the similarity values are below a threshold value for a given time period, the system will notify the operator so he doesn't need to face and deal with big data. With this method, the big data is summarized and abstracted into useful and applicable data sets where the information can be more easily extracted from. After that, for instance, this information can be used for predictive maintenance operations [11]. We choose Zuo et al.'s similarity measure of generalized fuzzy numbers for data comparison cause of its good deal with non-standard fuzzy numbers and fault detection [12]. In order to show the effectiveness and efficiency of our method in real-world situations, an IoT use case is presented and analyzed utilizing the method introduced in this paper.

This paper is structured as follows; Literature review is presented in the next section followed by the Methodology introduction. The case study is the fourth section followed by the conclusions.

II. LITERATURE REVIEW

Collecting useful information from big data analysis is an objective which requires scalable analytical methods to return well-timed results [5]. In the past, Agrawal and his co-workers proposed a method [13] for mining association rules from datasets utilizing quantitative and categorical attributes.

After that, the fuzzy set theory has been used more frequently in intelligent systems because of its simplicity, similarity to human reasoning and protection of main information from inaccuracy [14]. The Hong et al.'s generalized mining algorithm [9] integrates fuzzy-set concepts and generalized data mining technologies to find cross-level rules from quantitative data.

De Mauro et al. reviewed [5] some big data analytical methods that have been singled out by Chen et al. [15]. They presented a list of the most usual procedures that includes: cluster analysis, genetic algorithms, natural language processing, machine learning, neural networks, predictive modeling, regression models, social network analysis, sentiment analysis, signal processing, and data visualization [5].

The Chen et al. [7] did some researches in the field of analysis of time series and determined research directions including subsequence matching and full sequence matching. They presented that analysis of time series relies on three components, including representation, similarity measures, and indexing.

Marjani et al. [16] determined that big data analytics methods can be categorized as follows; classification, clustering, association rule mining, and prediction categories. Each category is a data summarization and mining function and involves many algorithms to fulfill information extraction and analysis requirements. For example, some methods offer classification such as Bayesian network, support vector machine (SVM), and k-nearest neighbor (KNN). Some other methods are prevalent in clustering such as partitioning, hierarchical clustering, and co-occurrence. Association rule mining and prediction include many significant methods as well.

This paper focuses on proposing a fuzzy data summarization and analytic method based on fuzzy similarity measure which is able to deal with quantitative data under a given time period.

III. RESEARCH METHOD

A. Generalized Fuzzy Number

As one of the fundamental concepts in this research, "generalized fuzzy number" is explained in this section. Let $\tilde{A}$ be a generalized fuzzy number. If $\tilde{A}$ is a generalized triangular fuzzy number then $\tilde{A} = (a_1, a_2, a_3; w_{\tilde{A}})$ and it is shown in Fig. 1. If $\tilde{A}$ is a generalized trapezoidal fuzzy number, then $\tilde{A} = (a_1, a_2, a_3, a_4; w_{\tilde{A}})$ and it is shown in Fig. 2 [17]:

Fig. 1. Generalized triangular fuzzy number

Fig. 2. Generalized trapezoidal fuzzy number

Compared to normal fuzzy numbers, the generalized fuzzy numbers can deal with uncertain information in a more flexible manner due to the use of parameter w which is the height of the fuzzy number and represents the degree of confidence of opinions of decision makers. According to de Barros et al. [18] the membership function ($\mu(x)$) of the fuzzy number number $\tilde{A}$ as a triangular fuzzy number (Fig. 1) can be presented as (1), and According to Farhadinia and Ban [19] the membership function ($\mu(x)$) of the fuzzy number $\tilde{A}$ as a trapezoidal fuzzy number (Fig. 2) can be presented as (2):

$$\mu_{\tilde{A}}(x) = \begin{cases} 0 & x < a_1 \\ \frac{w_{\tilde{A}}(x-a_1)}{a_2-a_1} & a_1 \leq x \leq a_2 \\ \frac{w_{\tilde{A}}(a_3-x)}{a_3-a_2} & a_2 \leq x \leq a_3 \\ 0 & x > a_3 \end{cases} \quad (1)$$

$$\tilde{A} = (a_1, a_2, a_3; w_{\tilde{A}}), a_1 \leq a_2 \leq a_3, 0 < w_{\tilde{A}} \leq 1$$

$$\mu_{\tilde{A}}(x) = \begin{cases} 0 & x < a_1 \\ \frac{w_{\tilde{A}}(x-a_1)}{a_2-a_1} & a_1 \leq x \leq a_2 \\ w_{\tilde{A}} & a_2 \leq x \leq a_3 \\ \frac{w_{\tilde{A}}(a_4-x)}{a_4-a_3} & a_3 \leq x \leq a_4 \\ 0 & x > a_4 \end{cases} \quad (2)$$

$$\tilde{A} = (a_1, a_2, a_3, a_4; w_{\tilde{A}}), a_1 \leq a_2 \leq a_3 \leq a_4, 0 < w_{\tilde{A}} \leq 1$$

B. Utilize Data to make Generalized Fuzzy Numbers

The big data generated by the IoT devices and sensors contains various components (e.g.: time, temperature, humidity, acceleration, coordination, air pressure, air quality, humidity, etc). In a time series, each component can be defined as a variable and shown by $x(i)$. So we can present our big data by vector x that consists of i number of variables and is shown in (3):

$$x = [x(1), x(2), \dots, x(i)] \quad (3)$$

In each case, a certain number of these variables are considered by the operator or decision maker hence he has a large amount of data about his considered variables to monitor, review and cleanse to be used for fault detection. we show the considered variables of decision maker by vector y that consists of j number of variables and is shown in (4):

$$y = [y(1), y(2), \dots, y(j)], j \leq i \quad (4)$$

In the method suggested by this paper, the vector of environmental factors (significant variables) is being represented by a generalized fuzzy number. If $j = 3$ which means that we have three significant variables then it will be represented by a triangular fuzzy number ($\tilde{A}$) which is presented in (5):

$$\tilde{A} = (y(1), y(2), y(3); w_{\tilde{A}}) = (a_1, a_2, a_3; w_{\tilde{A}}) \quad (5)$$

If $j = 4$ which means that we have four significant variables then it will be represented by a trapezoidal fuzzy number ($\tilde{A}$) which is presented in (6):

$$\tilde{A} = (y(1), y(2), y(3), y(4); w_{\tilde{A}}) = (a_1, a_2, a_3, a_4; w_{\tilde{A}}) \quad (6)$$

Note that depending on the unit utilized for each significant variable, variations should be comparable and normalized. Therefore, prior to the utilization of significant variables as a represented fuzzy number, a unit normalization for all significant variables should be performed. It means that each variables unit in the similarity analysis calculations must be selected in such a way that its numerical variations are proportional to the other variables.

C. Generalized Fuzzy Number Similarity Measure

In the case of IoT data analytics, there exist nonstandard fuzzy numbers, therefore, the similarity method must have the ability to measure the similarity of non-standard fuzzy numbers. Hence, we choose Zuo et al.'s similarity measure of generalized fuzzy numbers for our data comparison (and dealing with fault diagnosis). The calculation process for the similarity of two non-standard fuzzy number $\tilde{A}$ and $\tilde{B}$ is described in detail by Zuo et al. [12]. They utilize the following (7) for measuring the similarity of $\tilde{A}$ and $\tilde{B}$:

$$S(\tilde{A}, \tilde{B}) = S(\tilde{A}', \tilde{B}') = \left(1 - \sum_{i=1}^4 \frac{|a'_i - b'_i|}{4}\right) \left(1 - |x_{\tilde{A}'}^* - x_{\tilde{B}'}^*|\right) \times \frac{\min(P(\tilde{A}'), P(\tilde{B}')) + \min(a(\tilde{A}'), a(\tilde{B}'))}{\max(P(\tilde{A}'), P(\tilde{B}')) + \max(a(\tilde{A}'), a(\tilde{B}'))} \quad (7)$$

The x component of the center of gravity of $\tilde{A}'$ is represented by $x_{\tilde{A}'}^*$ and the The x component of the center of gravity of $\tilde{B}'$ is represented by $x_{\tilde{B}'}^*$. The perimeter of $\tilde{A}'$ is represented by $P(\tilde{A}')$ and the perimeter of $\tilde{B}'$ is represented by $P(\tilde{B}')$. The area of $\tilde{A}'$ is represented by $a(\tilde{A}')$ and the area of $\tilde{B}'$ is represented by $a(\tilde{B}')$.

Moreover, to measure the similarity between two triangular fuzzy numbers with the (7), we require to transform $\tilde{A}$ and $\tilde{B}$ to trapezoidal fuzzy numbers based on (1) and (2) with (8) and (9) as follows:

$$\tilde{A} = (a_1, a_2, a_3; w_{\tilde{A}}) \longrightarrow \tilde{A} = (a_1, a_2, a_2, a_3; w_{\tilde{A}}) \quad (8)$$

$$\tilde{B} = (b_1, b_2, b_3; w_{\tilde{B}}) \longrightarrow \tilde{B} = (b_1, b_2, b_2, b_3; w_{\tilde{B}}) \quad (9)$$

D. Data Analysis and Fault Detection

In the previous step, we used Zuo et al.'s similarity measure of generalized fuzzy numbers to compare IoT data with decision makers ideal reference point. It is very important that the ideal reference point of the decision maker is specified and can be presented as an ideal state with an intended and acceptable range for significant variables at a specified time interval. Accordingly, we can define the acceptable percentage of similarity for the system and determine when the system is out of acceptable range. In this way, the big data can be significantly summarized, and the operator only is required to check on the critical situations. The acceptable percentage of similarity and desired time interval depends on the acceptable data range of the decision maker. It can be calculated by computing the similarity percentage of borderline points.

IV. CASE STUDY

To show the significance of the method presented in this paper, we utilize a use case. In this use case, we deal with the employee's comfort in the workplace. Temperature, humidity, and lighting all play an important role in working environment comfort [20], [21]. We utilize sensors to collect data regarding three environmental factors of a workplace which are temperature, humidity, and lighting (it can be more

than three factors, for instance, air quality can be added as a fourth environmental factor). The ideal reference point for our analysis is set by standards describing a comfortable work places environmental factor (Table I). In this use case, the ideal reference point is compared with the recordings from the sensors, constantly to detect the anomalies on time. The ideal reference point for the temperature, humidity, and lighting is transformed to a triangular fuzzy number and is compared in real-time with each data point collected from the workplace which is also transformed to a triangular fuzzy number. The comparison illustrates the deviation from the ideal condition and assists through the fault detection. According to ASHRAE standard 55-2010 and AS 1680, we provide a summary of the ideal state and acceptable range for temperature, humidity, and lighting in Table I [22], [23]:

TABLE I
IDEAL WORKPLACE ENVIRONMENTAL FACTORS AND THEIR ACCEPTABLE RANGES [22], [23]

Environmental Factor	Ideal Status	Acceptable Range	Characteristics of the Activity and Interior
Temperature	Summer: 24.5°C Winter: 22°C	Summer: 23-26°C Winter: 20-23.5°C	Normal Office Work, PC Work, Study Library, Groceries, Show Rooms, Laboratories
Humidity	50%	40% - 60%	
Illuminance	500 Lux	400-600 Lux	

The sensor recordings are collected for a given time period by placing a sensor in a normal office room on an office desk. The objective is to constantly collect data so that the monitoring of the situation and detection of anomalies is possible. The data that has been gathered by sensors are presented in Appendix Table 2 and turned into vector x (with nine components) which is defined in (10):

$$x = [\text{Year, Month, Day, Hour, Minute, Second, Temperature, Humidity, Light}]$$

$$= [x(1), x(2), x(3), x(4), x(5), x(6), x(7), x(8), x(9)] \quad (10)$$

In this case, the significant variables for decision maker are temperature, humidity, and lighting so the significant variables vector (based on (4)) is shown below:

$$y = [x(7), x(8), x(9)]$$

The units of the recorded data are as follows; Celsius, humidity percentage, and lux, and there is a need to normalize the units for our method. According to the decision makers opinion, in this case, the approximate value of a one degree temperature change is equivalent to 10% humidity change and 100 lux lighting change. Therefore, the following unit changes are performed by (11) and (12):

$$x'(8) = x(8) \setminus 10 \quad (11)$$

$$x'(9) = x(9) \setminus 100 \quad (12)$$

Because we have three significant variables, so we utilize y to make triangular fuzzy number $\tilde{A}$ for each data set. The $\tilde{A}$ is presented in (13) as below:

$$\tilde{A} = (x(7), x'(8), x'(9); w_{\tilde{A}}) \quad (13)$$

Based on Table I, the ideal reference point can be shown by y_o as well as the triangular fuzzy number, $\tilde{B}$:

$$y_o = [24.5, 50, 500]$$

Transformation to fuzzy number tilde B

$$\tilde{B} = (24.5, 5, 5; 1)$$

Note that we consider the height of all fuzzy numbers equal to one in this case study.

Based on Table 1 and Zuo et al.'s similarity measure (7), we calculated the acceptable similarity to ideal status for the data which is above 92% and can be changed based on decision makers preferences. Moreover, the length of each anomaly period is set to be 15 minutes before the alarm is being set off (informing the facility owner or manager) in other words if the similarity of the recorded data from the ideal reference point is out of range for 15 minutes, the fault has occurred and the system must be checked by an operator (Fig. 3). If the recorded anomaly period is less than 15 minutes, in this case, it is considered as a noise. Depending on the use case (e.g.: a power plant may require much tighter control cycles), out of range period can be different and should be set by the decision maker.

Fig. 3. The similarity of workplace conditions (sensory data) to the ideal state over a period of 1.5 hrs.

As illustrated in Fig. 3, between 20 : 52' : 58'' and 21 : 47' : 09'', the environmental factors of the office room are in the acceptable range (normal state) as its similarity to the ideal state is higher than 92%. At 21 : 48' : 39'' the similarity to ideal state decreases to 89.91% and it stays under the control line of 92% similarity until 21 : 54' : 41''. As is seen in Appendix, in the time interval from 21 : 48' : 39'' to 21 : 54' : 41'', the lighting has decreased by almost 220 lux while the temperature and humidity are in the acceptable range. This anomaly period is less than 15 minutes, so it is considered as a noise.

Between 21 : 56' : 11'' and 22 : 03' : 43'', the environmental factors of the office room are back in the acceptable

range as its similarity to the ideal state goes over than 92%. At 22 : 05' : 13" the similarity to ideal state decreases to 89.61% and it stays under the control line of 92% similarity until 22 : 23' : 17". As is seen in Appendix, in the time interval from 22 : 05' : 13" to 22 : 23' : 17", the lighting has decreased by almost 220 lux while the temperature and humidity are in the acceptable range. This time the anomaly period is more than 15 minutes so the operator must be informed to check the office rooms lighting system.

V. CONCLUSIONS

In this paper, we have proposed a new method of summarization and fault detection of big data based on a fuzzy similarity measure. Also, the usefulness and applicability of this method are being illustrated in an IoT use case. In other words, the proposed method is tested in a use case for the monitoring of environmental factors in a workplace. The proposed method efficiently summarizes the data so that the operator avoids dealing with the big data and only interfere when it is necessary. As the case study illustrates, the implementation of the proposed method frees an operator from monitoring different environmental factors individually, while providing a holistic view of the system in a single similarity measurement value. Moreover, when the similarity value goes out of acceptable range for a given time period, the operator is alarmed and can dig deeper to identify the root cause. Compared to conventional data summarization methods for quantitative data, our approach gets smoother summarizing results due to its fuzzy characteristics. And, our proposed method takes less time to fuzzification and defuzzification, due to its simplicity.

Although the proposed method works well for data mining and data analysis of big data, there is still much work needs to be done to expand this method to analyze more than four variables in fuzzy number similarity measurement. In other words, our method is currently capable of comparing a maximum of four different variables at a time due to the fuzzy numbers characteristics. Therefore, as an area for future researches, we propose the development of a method that can compare more individual variables simultaneously that our proposed method. Also, we used Zuo et al.'s similarity method that can be replaced by another efficient fuzzy similarity method to enable non-standard fuzzy numbers similarity measurement.

REFERENCES

- [1] M. Hung, "Leading the iot, gartner insights on how to lead in a connected world," *Gartner Research*, pp. 1–29, 2017.
- [2] S. D. T. Kelly, N. K. Suryadevara, and S. C. Mukhopadhyay, "Towards the implementation of iot for environmental condition monitoring in homes," *IEEE Sensors Journal*, vol. 13, no. 10, pp. 3846–3853, 2013.
- [3] M. Farooq, M. Waseem, S. Mazhar, A. Khairi, and T. Kamal, "A review on internet of things (iot)," *International Journal of Computer Applications*, vol. 113, no. 1, pp. 1–7, 2015.
- [4] S. Madakam, R. Ramaswamy, and S. Tripathi, "Internet of things (iot): A literature review," *Journal of Computer and Communications*, vol. 3, no. 05, p. 164, 2015.
- [5] A. De Mauro, M. Greco, and M. Grimaldi, "A formal definition of big data based on its essential features," *Library Review*, vol. 65, no. 3, pp. 122–135, 2016.
- [6] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile networks and applications*, vol. 19, no. 2, pp. 171–209, 2014.
- [7] F. Chen, P. Deng, J. Wan, D. Zhang, A. V. Vasilakos, and X. Rong, "Data mining for the internet of things: literature review and challenges," *International Journal of Distributed Sensor Networks*, vol. 11, no. 8, p. 431047, 2015.
- [8] C. Zhang, G. Zeng, and X. T. HongboWang, "Analysis on data mining model objected to internet of things," *International Journal of Advancements in Computing Technology*, vol. 4, no. 21, 2012.
- [9] T.-P. Hong, K.-Y. Lin, and S.-L. Wang, "Fuzzy data mining for interesting generalized association rules," *Fuzzy sets and systems*, vol. 138, no. 2, pp. 255–269, 2003.
- [10] G. Florez, S. M. Bridges, and R. B. Vaughn, "An improved algorithm for fuzzy data mining for intrusion detection," in *Annual Meeting of The North American Fuzzy Information Processing Society Proceedings*, vol. 5, 2002.
- [11] S. H. Khajavi, J. Holmström, and M. Baumers, "Additive manufacturing as a platform for introducing cyber-physical services," in *Proceedings of the 3rd International Conference on Progress in Additive Manufacturing (Pro-AM 2018)*, 2018, pp. 637–643.
- [12] X. Zuo, L. Wang, and Y. Yue, "A new similarity measure of generalized trapezoidal fuzzy numbers and its application on rotor fault diagnosis," *Mathematical Problems in Engineering*, vol. 2013, 2013.
- [13] R. Srikant and R. Agrawal, "Mining quantitative association rules in large relational tables," in *Acm Sigmod Record*, vol. 25, no. 2. ACM, 1996, pp. 1–12.
- [14] A. Kandel and G. Langholz, *Hybrid architectures for intelligent systems*. CRC press, 1992.
- [15] H. Chen, R. H. Chiang, and V. C. Storey, "Business intelligence and analytics: from big data to big impact," *MIS quarterly*, pp. 1165–1188, 2012.
- [16] M. Marjani, F. Nasaruddin, A. Gani, A. Karim, I. A. T. Hashem, A. Siddiqui, and I. Yaqoob, "Big iot data analytics: architecture, opportunities, and open research challenges," *IEEE Access*, vol. 5, pp. 5247–5261, 2017.
- [17] C. P. Pappis and N. I. Karacapilidis, "A comparative assessment of measures of similarity of fuzzy values," *Fuzzy sets and systems*, vol. 56, no. 2, pp. 171–174, 1993.
- [18] L. C. De Barros, R. C. Bassanezi, and W. A. Lodwick, *First Course in Fuzzy Logic, Fuzzy Dynamical Systems, and Biomathematics*. Springer, 2016.
- [19] B. Farhadinia and A. I. Ban, "Developing new similarity measures of generalized intuitionistic fuzzy numbers and generalized interval-valued fuzzy numbers from similarity measures of generalized fuzzy numbers," *Mathematical and Computer Modelling*, vol. 57, no. 3–4, pp. 812–825, 2013.
- [20] W. Victoria, "Officewisea guide to health & safety in the office (edition no. 5, jan 2006)," *WorkSafe Victoria, Melbourne*, 2006.
- [21] A. J. SUMMERS, "Lighting and the office environment: A review," *Journal of Physiotherapy*, vol. 35, no. 1, pp. 15–24, 1989.
- [22] W. Victoria, "Compliance code: workplace amenities and work environment (edition no. 01, sept 2008)," *WorkSafe Victoria, Melbourne*, 2008.
- [23] F. M. Mele, "A model-based approach to hvac fault detection and diagnosis," 2012.

APPENDIX

TABLE II
THE COLLECTED DATA BY SENSORS IN 2018-08-29

Data Sets	Time	Temperature (°C)	Humidity (%)	Lighting (lux)	Similarity to Ideal Status (%)
No.	x(4):x(5):x(6)	x(7)	x(8)	x(9)	
1	20: 52: 58	24.3	42	452	96.04
2	20: 54: 28	24.3	42	456	96.16
3	20: 55: 59	24.3	42	458	96.22
4	20: 57: 29	24.3	42	458	96.22
5	20: 58: 59	24.3	42	458	96.22
6	21: 00: 29	24.2	42	457	96.33
7	21: 02: 00	24.2	42	458	96.36
8	21: 03: 30	24.2	42	457	96.33
9	21: 05: 00	24.2	42	457	96.33
10	21: 06: 31	24.2	42	457	96.33
11	21: 08: 01	24.2	42	457	96.33
12	21: 09: 31	24.2	42	457	96.33
13	21: 11: 01	24.2	42	458	96.36
14	21: 12: 32	24.2	42	454	96.24
15	21: 14: 02	24.2	42	453	96.21
16	21: 15: 32	24.2	42	453	96.21
17	21: 17: 03	24.2	42	454	96.24
18	21: 18: 33	24.2	42	453	96.21
19	21: 20: 03	24.2	42	458	96.36
20	21: 21: 34	24.2	42	458	96.36
21	21: 23: 04	24.2	42	458	96.36
22	21: 24: 34	24.2	42	459	96.39
23	21: 26: 05	24.2	42	458	96.36
24	21: 27: 35	24.1	42	459	96.54
25	21: 29: 05	24.1	42	458	96.51
26	21: 30: 36	24.1	42	458	96.51
27	21: 32: 06	24.1	42	458	96.51
28	21: 33: 36	24.1	42	459	96.54
29	21: 35: 07	24.1	42	459	96.54
30	21: 36: 37	24.1	42	459	96.54
31	21: 38: 07	24.1	42	460	96.57
32	21: 39: 37	24.1	42	456	96.44
33	21: 41: 08	24.1	42	455	96.41
34	21: 42: 38	24.1	42	456	96.51
35	21: 44: 08	24.1	42	455	96.49
36	21: 45: 39	24.1	42	455	96.41
37	21: 47: 09	24.0	42	456	96.51
38	21: 48: 39	24.0	42	236	89.91
39	21: 50: 10	24.0	42	232	89.92
40	21: 51: 40	24.0	42	232	89.93
41	21: 53: 10	24.0	42	231	89.94
42	21: 54: 41	24.0	42	232	89.95
43	21: 56: 11	24.0	42	442	96.15
44	21: 57: 41	24.0	42	453	96.46
45	21: 59: 12	24.0	42	455	96.49
46	22: 00: 42	24.0	42	455	96.49
47	22: 02: 12	24.0	42	454	96.47
48	22: 03: 43	24.0	42	452	96.44
49	22: 05: 13	24.0	42	223	89.61
50	22: 06: 43	23.9	42	224	89.62
51	22: 08: 13	23.9	42	224	89.63
52	22: 09: 44	23.9	42	224	89.64
53	22: 11: 14	23.9	42	223	89.65
54	22: 12: 44	23.9	42	224	89.66
55	22: 14: 15	23.9	42	224	89.67
56	22: 15: 45	23.9	42	224	89.68
57	22: 17: 15	23.9	42	224	89.69
58	22: 18: 46	23.9	42	224	89.70
59	22: 20: 16	23.8	42	224	89.71
60	22: 21: 46	23.8	42	224	89.85
61	22: 23: 17	23.8	43	224	90.15