TRANSFORMING AND PROCESSING THE MEASUREMENT SIGNALS

DATA CLEANING METHOD IN WIRELESS SENSOR-BASED ON INTELLIGENCE TECHNOLOGY

Roman Diachok, PhD Student, Halyna Klym, Dr.Sc., Prof.,
Lviv Polytechnic National University, Ukraine; e-mail: rodyachok@gmail.com, halyna.i.klym@lpnu.ua

Abstract. The method of cleaning management data in wireless sensor networks based on intelligence technology has been studied. Specific forms of application of wireless sensor networks are analyzed. The characteristics of the structure of wireless sensor networks are presented and the data cleaning technology based on the clustering model is offered. An algorithm for deleting a cluster-based replication record is proposed and the accuracy of data cleaning methods is tested. The obtained results testify to the efficiency of using the studied method.
Key words. Wireless sensor, Network control, Database cleaning.

1. Introduction

The rapid development of the Internet and the popularity of computers marked the entry of mankind into the era of online information [1]. This was facilitated by the rapid development of the World Wide Web, which initiated the exponential growth of online information [2]. It is known that the sources of information are quite extensive. The largest sources of data are network information. In addition to relational databases, distributed databases, etc., they have all achieved considerable development [3]. The amount of data available is growing. But at the same time, one problem after another is emerging, the most common of which is data integration. The latter is an important step in information processing in many areas [4]. For a large set of data, the criterion for the quality of the relevant parameters is the accuracy of the data obtained in the integration process.

2. Drawbacks

For a large data set, the criterion is the quality of the relevant parameters, such as the quality and accuracy of the data in the integration process to assess whether the data integration is excellent [5]. However, some mistakes cannot be avoided when integrating the process. In general, the main reason for the problems in the integration process is that there are no harmonized standards among the databases, and the data format is different, which affects the data. Integration has caused some obstacles [6]. Therefore, when entering large amounts of data, there will always be some errors and conflicting results [7].

It is known that a wireless sensor network is a technology for receiving and processing information, which mainly consists of sensors, MEMS, and network systems. Compared to the traditional era of the PC, it is characterized by smaller size, lower price, and goes beyond the traditional computer 8]. Each sensor unit can measure and analyze signals in the environment with the help of built-in multiple sensors and obtain the necessary data [9]. Compared to a traditional network, a wireless sensor network focuses on the data received, not on its transmission. The wireless sensor has a wide range of functions that can not only monitor the environment, and the condition of the building, but also control a smart home with certain technical means [10]. The use of sensor sensors for military purposes is another important focus of the study of the sensor network. Given the main characteristics of wireless sensor networks, this paper conducted an in-depth study of data cleaning methods and talent management in wireless sensor networks based on intelligence technology. With this method, talent management data in wireless sensor networks can be implemented independently. If it is effective, the total amount of data decreases. Reducing talent management data can not only increase the efficiency of the analysis process but also improve the quality of analysis results.

3. Aim of the work

This work aims to analyze the cleaning methods while repeated data recording as well as perform an in-depth study of data cleaning methods in wireless sensor networks based on intelligence technology for decreasing the scale of data management and enhancing the data analysis efficiency and the quality.

4. Methodology of Issue

Wireless sensor networks mainly consist of sensor nodes, detection zones, and servers. Sensor nodes can be located near objects that require measurement data using manual deployment. Once deployed, these sensory nodes self-organize in some way and begin to share the environment and objects to obtain the necessary data. This self-organizing form can form a corresponding network and transmit all data back to the main node via the relay mode. Eventually, all data in the node is transmitted to the server through the communication system. When users use wireless sensors to obtain data, it becomes possible to efficiently collect the necessary data through the management and control of nodes. The architecture of a typical wireless sensor network [1] is shown in Fig. 1.

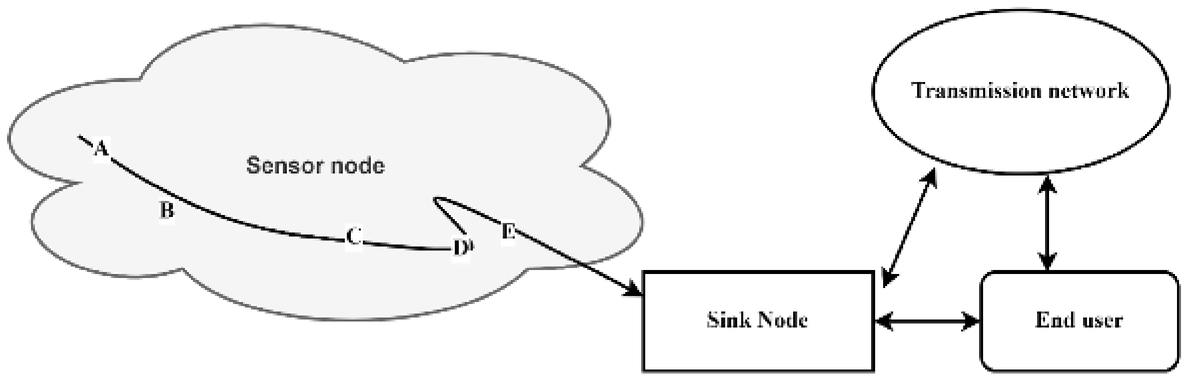

Fig. 1. The topological structure of typical sensor networks

The sensor unit consists of a sensor module, a processing module, a wireless communication module, and a power supply module. The sensor module is responsible for receiving information and converting data. The processing module monitors the operation of the entire sensor node, processes the data collected by him and the data sent by other nodes, then launches a network protocol to control the process of the communication node; the wireless communication module is especially connected to other sensor nodes of sending and receiving data; the power supply module provides power to the sensor units. For network operation, each node sensor in a wireless sensor must consider both traditional network nodes and routers, not only to collect and process local information but also to process data transmitted from other nodes. In the process of data transfer, nodes must be able to work together. At this stage, the hardware and software technology of the sensor node is the focus of sensor network research, as it has a powerful ability to process, store and transmit node data. By connecting the sensor to the Internet, you can convert communication between different network protocols. At the same time, the sensor can distribute tasks to all nodes simultaneously and transmit the collected data to an external network. For different applications, the composition of the sensors is also different, but almost all sensors have common characteristics. These typically include a sensor unit, a processing unit, a wireless unit, a power supply, and a traditional data sensor [4]. The relationship of the components is shown in Fig. 2.

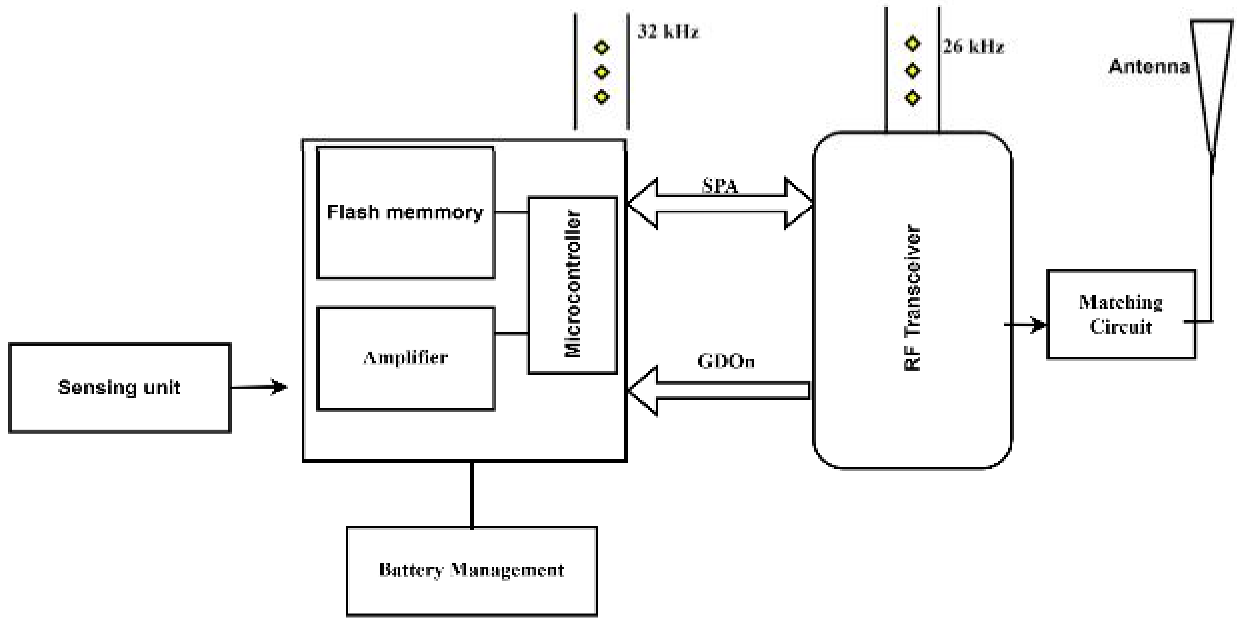

Fig. 2. Sensor node components

The composition of the sensor unit is usually relatively simple and generally consists of a sensor and a functional module of analog-to-digital conversion, which is responsible for converting data to obtain information in the observation area. The main part of the processor unit is the built-in system, which includes the CPU, and memory, and is generally responsible for managing the nodes of the entire sensor and storing the collected data, as well as processing data received by other nodes. The main function of a wireless device is to complete the data transfer without using a wired device. The main part of the power supply unit is the power supply module, the main function of which is to provide energy to the sensor units. There are also some other modules such as positioning systems and mobile systems. Thanks to the cooperation of these devices, the wireless sensor network can work properly. When using a wireless sensor network to obtain data, to make it more accurate and efficient, you need to configure a large number of sensor nodes. Therefore, the number of sensor nodes can be very large. Because the number of sensor units is relatively large and the volume is small, and staff in some areas may not arrive on time, the sensors cannot be supplemented by battery replacement. At this point, it makes sense to calculate the power consumption of the sensor units. The main energy-intensive part of the sensor unit is very energy-intensive when transmitting data to the wireless module. The wireless module has four wireless states: send, receive, standby, and sleep. The relationship between wireless power consumption and distance is shown in formula 1:

$$E = kd^n, \tag{1}$$

here E is the power consumption of the wireless communication network, d is the distance, d and k are the constants.

As the communication distance increases, the energy consumption will increase sharply.

The mathematical model of Bayesian data network analysis, which is commonly used in the process of data intelligence, is presented in formula 2:

$$P(x_1,\ldots x_n) = P(x_1)P(x_2/x_1)P(x_n/x_1,\ldots x_{n-1}) \tag{2}$$

The formula for the degree of confidence of the model is given in expression (3):

$$p(Mi/D) = \frac{p(D/Mi)p(Mi)}{p(D)}, \tag{3}$$

here p is the probability of displaying the edge. Bayesian information criterion (BIC) is a large selective approximation of the probability of an edge. Using the Laplace approximation, a large sample approximation can be performed for P, and the ICD estimation function can be derived, then the logarithmic likelihood function can be extended by estimating the maximum likelihood, and then the calculation can be transformed into a multidimensional normal distribution function. First, the Laplace approximation is used for the a posteriori probability, as shown in equation (4):

$$p(D/m) = \int p(D/\theta, m)p(\theta/m)d\theta \tag{4}$$

Due to the definition of model data, it becomes convenient to use data analysis technology for their processing. Current management data faces the problem of sudden data increase. These large databases usually contain errors or inconsistencies for some reason. Causes of errors include incorrect input, which leads to incorrect values because the entered data is inconsistent due to different formats or the use of different abbreviations can not fully collect information about the data and lead to loss of data. To avoid this situation, they are trying to solve the so-called "garbage disposal" problem. The process of clearing data is to resolve common errors and inconsistencies in large databases. Slight simple pre-processing before data cleaning can improve the quality of data cleaning in general. The block diagram of data cleaning is shown in Fig. 3.

The main process of data processing pre-processing is shown in Table 1.

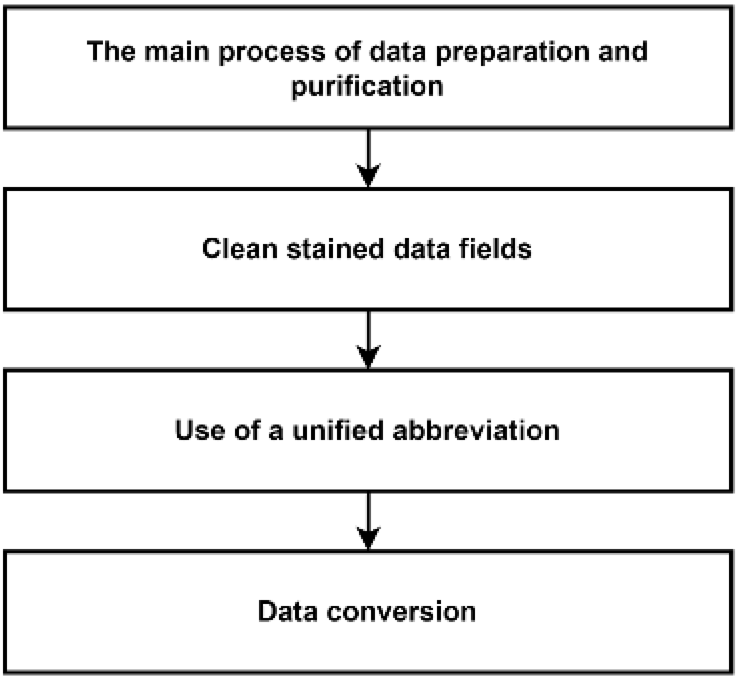

Fig. 3. The main process of cleaning and pre-processing

Table 1

The main process of cleaning and pre-processing

Stages of processing	A detailed description of the stages of data processing
Clean stained data fields	The main purpose of this step is to eliminate data entry errors. Some simple errors when correcting data records with some external features and external source files, such as checking that the city's zip code matches and that the date of birth and age match. This will increase the accuracy and standardization of the data, as well as effectively avoid the clustering process, as the data error is too large to record the same object in more than one cluster.
Use of a unified abbreviation	According to the appropriate ratio of abbreviation and full name, all data is processed in a standardized way or unified abbreviation or representation of the full name.
Data conversion	In this process, we convert some data into various formats. In the database, the man is represented in the database "a", and in another database expressed "1", which creates conflicting data. The process of transforming data is to transform this conflicting data into consistent data. This process can also transform a data table into a data table with many different structures according to certain requirements.

When merging large databases, problems are often encountered: incorrect data entry, different schemes, or inconsistencies in the form of abbreviations. These problems result in a merged database to obtain multiple records that represent the same entity, but a slightly different attribute of values, thus creating conflicting results. After cleaning and pre-processing, some simple errors in the database are cleared. However, the object to be processed is a large database. The amount of data that needs to be processed is very large, so it leads to many errors and questionable results. The accuracy index used in this paper is a pure clustering comparison. The definition of pure clustering refers to the fact that all records contained in the cluster represent the same entity. The experimental method is used to evaluate data in a large-scale database. The goal of accuracy in the measurement process is the entire database, not just one data in the database. When data is recorded using pure clustering, the representation of the records is the same. If the records have different forms, this form of clustering is not pure clustering, which indicates that the clustering method is inaccurate. Using a cluster-based replication delete algorithm can significantly solve the problem of data mismatch. This method allows you to reduce the amount of data processing and increase the efficiency of data processing.

5. Consideration of Obtained Results

To make the experiment more exact and to be able to effectively verify the accuracy and efficiency of the algorithm, the clustering data that have been used for analysis are known, and specific values of the applicable data are determined. The data processed in the experiment is talent management record data. The attributes of the record included seven attributes of talent management. The experiment includes 1075 records. Due to some processing of copies, and then application of the method of processing random 547 errors, get the total number of effective records. Number - 2412. A total of 318 clusters containing more than two records are calculated manually, of which the largest cluster contained 3521, a total of 24 records.

Canopy clustering detection technology is applied to detect duplicate records. There are three main detection parameters: the distance thresholds T1 and T2 and the constant-coefficient *k*. The choice of T1 and T2 determines the size of the Canopy and the degree of its overlap, ie the amount of data that must be accurately calculated. Selecting the value of k determines whether the records can be accurately grouped. At the beginning of data processing, it is necessary to create values T1 and T2. The method of inverted detection of two values is considered in their work to establish them. In the case of different T1 and T2 (T1≤T2), the system clustering must calculate the calculated number of appropriate pairs to measure the quality of T1 and T2. The calculation of the different values of T1 and T2 is shown in Table 2.

Table 2

Calculation of different values of T1 and T2

T1 / T2	0.95	0.96	0.97	0.98	0.99
0.95	2167	6732	8786	11.374	12.596
0.85	-	2143	2653	8678	11.876
0.75	-	-	2114	2988	10.905
0.65	-	-	-	2187	10.245
0.55	-	-	-	-	10.256

Table 3

The clustering coefficient of different values

k	6	4	3	2	1
Clustering factor	0.832	0.88	0.988	1.155	1.46

According to the experimental data shown in Table 2, at T1 = 0.75 and T2 = 0.75 data points that need to be accurately calculated are the smallest.

Therefore, T1 = 0.75 and T2 = 0.75 are selected, which means that Canopy does not overlap. According to the ratio of the number of clusters obtained and the actual clustering, the value of k is shown in Table 3.

Table 3 shows that at k = 3 the clustering factor is closest to the real clustering, so the experiment chooses the distance threshold k = 3. However, the clustering factor still does not reach 1, because a random data error makes some data records incorrect by classifying them into one cluster. The home address of the talent management record table is a composite attribute. The program's data conversion method decomposes different parameters into sub-attributes and at the same time performs pre-processing before cleaning the data based on the use of an external source file. If there is a duplicate parameter corresponding to the existing ones then such data can be cleared they are considered dirty data. The amount of experimental data is quite small, covering only Guangxi, so setting 100% accuracy is relatively easy when setting up external source files.

Figure 4 demonstrates the results of experiments comparing pre-processed and unprocessed test copy recording methods, including the adjacent sort method and the test copy recording method by Canopy technology, in which the window size of the sort neighbor method selects two to compare the situation. Fig. 4 visualizes that the accuracy of the detection method of the pre-processed duplicate record is higher than the accuracy of the untreated duplicate record of the detection method. However, because the experimental data used are not large, the pre-processing cannot be fully reflected. In two cases ω = 24 and ω = 16, ω chooses 24 more accurately than 16, because the largest cluster in the experimental data contains 15 records. If ω = 16, this will lead to some duplicate records that cannot be detected, although ω = 16 when you need to make many unnecessary comparisons, which increases the number of calculations, which can guarantee a higher level of accuracy. In this experiment, the pre-treated neighbor ranking method ω = 24 is the same as the Canopy cluster replication record detection method, but the Canopy method has a higher response rate, indicating that it can obtain more replication records. From so the algorithm is more effective.

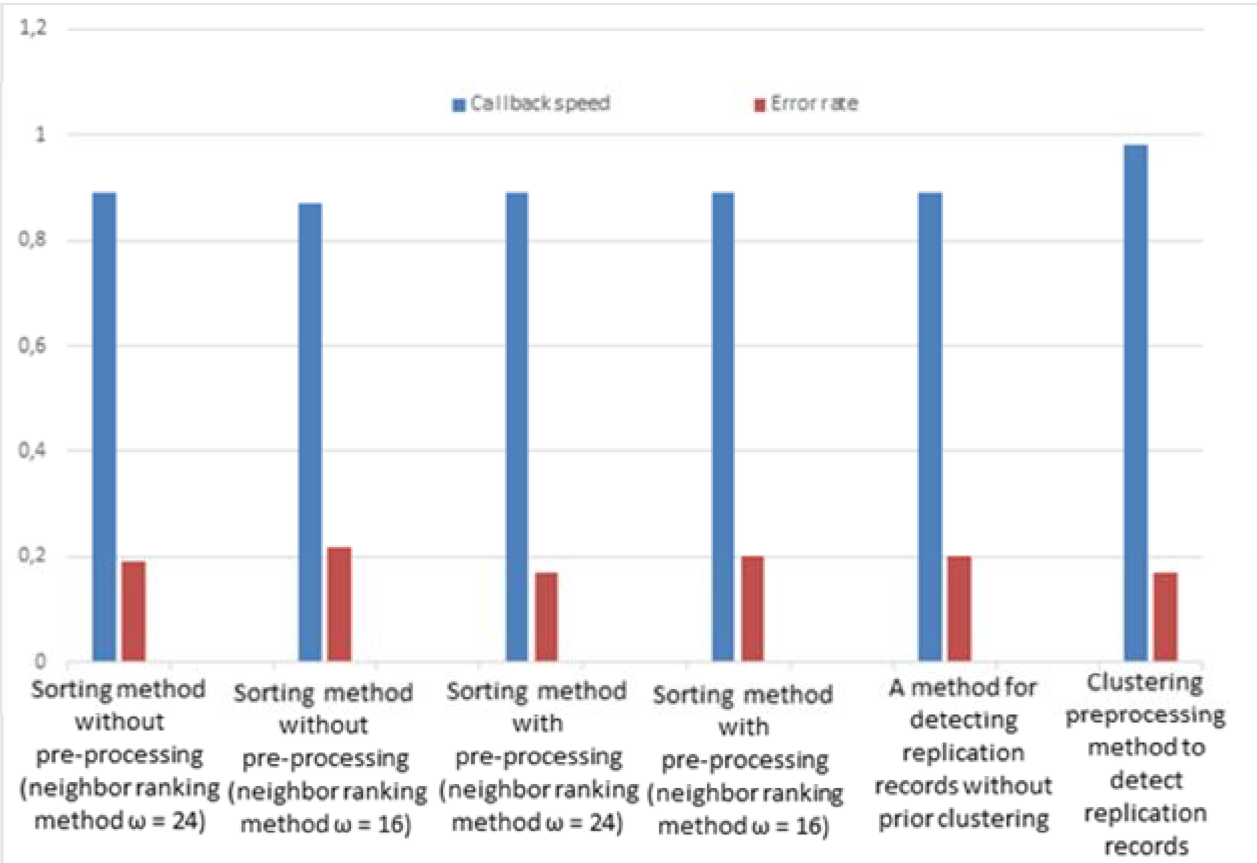

Fig. 4. The results of the comparison of the preliminary processing of the test

6. Conclusions

The method of data cleaning is an extremely important research area, which allows for resolving data inconsistencies in the identification of the same object and increases the accuracy of recognition. Due to faster and more convenient access to information, the amount of data is increasing every day. While analyzing data and making business decisions, it needs to combine a few data information to simplify searching for the correct model. The wrong choice inevitably leads to emerging the incorrect or conflicting data. The merge results in the approximation of duplicate entries, which is prohibited in the database, and these duplicate entries must be deleted.

7. Gratitude

The authors express their gratitude to the Ministry of Education and Science of Ukraine for the support thanks to Project No 0122U000807 for young scientists.

8. Conflict of Interest

The authors state that there are no financial or other potential conflicts regarding this work.

References

[1] L. Zhu, M. Li, Z. Zhang, et al., Big data mining of users' energy consumption patterns in the wireless smart grid [J]. IEEE Wirel. Commun. 25(1), 84–89 (2018), DOI:10.1109/MWC.2018.1700157,
https://ieeexplore.ieee.org/document/8304397

[2] W. Sun, L. Zhang, Y. Zhang, et al., Enhanced works of separation for (00 01)ZnO|(111)ZrO_2 interfaces via ion-doping in ZnO: Data-mining and density functional theory study [J]. Comput. Mater. Sci. 142, 410–416 (2018), DOI:10.1016/j.commatsci.2017.10.044
https://www.sciencedirect.com/science/article/abs/pii/S0927025617306080

[3] S. Shafiee, S. Minaei, Combined data mining/NIR spectroscopy for purity assessment of lime juice [J]. Infrared Phys. Technol. 91, 193–199 (2018), DOI:10.1016/j.infrared.2018.04.012 https://www.sciencedirect.com/science/article/abs/ pii/S135044951730662X

[4] D.W. Upton, B.I. Saeed, P.J. Mather, et al., Wireless sensor network for radiometric detection and assessment of partial discharge in high-voltage equipment [J]. Radio Sci. 53(3), 357–364 (2018), DOI:10.1002/2017RS006507
https://ieeexplore.ieee.org/abstract/document/8679774

[5] P.V. Mekikis, E. Kartsakli, A. Antonopoulos, et al., Connectivity analysis in clustered wireless sensor networks powered by solar energy [J]. IEEE Trans. Wirel. Commun. 17(4), 2389–2401 (2018), DOI: 10.1109/TWC.2018.2794963
https://ieeexplore.ieee.org/abstract/document/8267240

[6] D. Aygör, S.U. Rehman, F.V. Çelebİ. Impact of buffer management solutions on MAC Layer Performance in Wireless Sensor Networks. IEICE Transac. Commun. E101.B(9), 2058–2068 (2018), DOI: 10.1587/transcom.2017EBP3389
https://www.jstage.jst.go.jp/article/transcom/advpub/0/advpub_2017EBP3389/_article/-char/ja/

[7] A. Alomari, F. Comeau, W. Phillips, et al., New path planning model for mobile anchor-assisted localization in wireless sensor networks [J]. Wirel. Netw 8, 1–19 (2018), DOI:10.1088/1742-6596/1176/2/022003
https://link.springer.com/article/10.1007/s11276-017-1493-2

[8] L. Kumar, V. Sharma, A. Singh, Cluster-based single-sink wireless sensor networks and passive optical network converged network incorporating sideband modulation schemes [J]. Opt. Eng. 57(2), 1 (2018). DOI:10.1117/1.OE.57.2.026102
https://www.spiedigitallibrary.org/journals/optical-engineering/volume-57/issue-2/026102/Cluster-based-single-sink-wireless-sensor-networks-and-passive-optical/10.1117/1.OE.57.2.026102.full

[9] W.K. Lee, M.J.W. Schubert, B.Y. Ooi, et al., Multi-source energy harvesting and storage for floating wireless sensor network nodes with long range communication capability [J]. IEEE Trans. Ind. Appl. 54(3), 2606–2615 (2018)
DOI: 10.1109/TIA.2018.2799158
https://ieeexplore.ieee.org/document/8272444

[10] W. Zhang, J. Yang, Y. Fang, et al., Analytical fuzzy approach to biological data analysis [J]. Saudi J. Biol. Sci. 24(3), 563–573 (2017). DOI: https://doi.org/10.1016/j.sjbs.2017.01.027
https://www.sciencedirect.com/science/article/pii/S1319562X17300360?via%3Dihub