Computing and Informatics, Vol. 42, 2023, 1136–1157, doi: 10.31577/cai_2023_5_1136

CLUSTERING MINING ALGORITHM OF INTERNET OF THINGS DATABASE BASED ON PYTHON LANGUAGE

Fang Wan*

Business College, Nanchang Jiaotong Institute
Nanchang 330100, China
e-mail: ppap0112@163.com

Ying Liu

School of Artificial Intelligence, Nanchang Jiaotong Institute
Nanchang 330100, China
e-mail: liuying10100160@163.com

Abstract. In order to solve the problems of reading delay in data mining of the Internet of Things database, a clustering mining algorithm of the Internet of Things database based on Python language is proposed. We designed an improved crawler algorithm based on the open-source structure of scratch through Python language, judge the similarity of recruitment data topics in the Internet of Things database through Bayesian classifier, and crawl the recruitment data in the Internet of Things database: the number of keywords in the text space, the degree of keyword extraction, and the number of keyword data in the text space. The time series model is used to eliminate the delay of text features. On this basis, the semi-supervised learning and semi-cluster analysis method is used to construct the corresponding classifier, complete the adaptive classification process of the text data stream and realize the clustering mining of the Internet of Things database based on Python language. The experimental results show that this method has a low reading delay, and can mine the attention, number of posts and click time frequency of the Internet of Things database from which the recruitment data are obtained.

Keywords: Python language, Internet of Things database, data clustering, data mining

* Corresponding author

1 INTRODUCTION

The technical field of the Internet of Things is developing very fast, and there are many places to use [1]. There is more and more research on its related intelligent devices, and the type and quantity of products are also being used. Therefore, the application of this system or device has been quite common in people's daily lives. The rapid development of the Internet has driven the progress of big data. We can find a kind of useful data in a large amount of data. This mining method has brought great help to people and has become one of the aspects that many people are willing to explore and develop. Among them, Python is a popular program language in this field, because it has a lot of content, powerful technology and powerful computing power [2].

In reference [3], aiming at the problem of low clustering efficiency of moving objects in convergent pattern mining of the Internet of Things, a spatiotemporal feature mining algorithm based on pattern growth and multi-minimum support is proposed. Mining frequent and asynchronous periodic spatiotemporal motion patterns, modeling the position sequence and adding the time information to the model. Then the sequential patterns of asynchronous cycles are deeply and recursively mined to realize data clustering mining. Clustering methods can be applied to a wide variety of database systems. The information can take the form of quantitative, category, or interval-based information. Reference [4] proposed a two-level distributed clustering routing algorithm based on unequal clusters. The main idea is to reduce the data transmission distance between member nodes and cluster heads and alleviate the hot spot problem by distributing two cluster heads in each cluster, so as to achieve energy saving and load balancing. The sensor nodes are clustered by the clustering method. A further statistic frequently utilized to assess the effectiveness of clustering techniques is cooperative knowledge. It measures how distinct two descriptions for identical information are. The clustered sensor nodes can transmit data to the receiver through different paths with minimum energy consumption. Reference [5] proposed the first memetic algorithm to solve the balanced classical least square sum clustering problem. The algorithm combines responsive threshold search and a backbone-based crossover operator to generate offspring and realizes data clustering mining. However, in the process of practical application, the above methods are prone to problems such as long response time due to the influence of data type and data scale in the Internet of Things database.

Python can develop languages for different objects. It is very powerful in computing and has a wide range of code resources. Therefore, Python has gradually become a tool used in the direction of data mining, and more and more tools are used for development [6]. Using Python to conduct this kind of work is not only convenient and requires less, but also involves popular relevant tool libraries, which will greatly reduce the steps of each link in the work, which will save a lot of time for researchers to focus on the design and development of data mining direction, so as to get better and more accurate results [7]. The foundation of this method is machine learning. It arranges the parameters or elements in an information source into

predetermined types or groups. Methods for data mining include synthetic neural networks, selection orchards, analytics, linear regression, and more. Therefore, this paper studies the clustering mining algorithm of the Internet of Things database based on Python language.

2 CONSTRUCTION OF CLUSTERING MINING ALGORITHM FOR INTERNET OF THINGS DATABASE BASED ON PYTHON LANGUAGE

Among the technologies for processing big data in the existing Internet platforms, many have not been optimized enough in terms of technical activity, and the reflection effect from users is not good. At the same time, there is no such technology that can meet the whole platform operated by users in the market. The current service platforms have relatively high prices in the market and lack the framework of technology types suitable for small companies or scientific research. Contract administration services, e-signature techniques, online marketing, consumer connection administration (CRM), payroll systems, interaction as well as collaborative techniques, as well as access control techniques, are just a few examples. In the past, private organizations relied on newspaper advertisements or posters at the end of the road. They now have access to considerably more potent tools for expanding their firm. Python has now become a well-developed and most frequently used program in the database field [8, 9]. Because it has very rich resources and powerful computing functions, it can help to complete the work more effectively in the mining process. To access greater reserves, underwater miners are more costly. Shallower as well as less valued resources are often mined using subsurface methods. Therefore, this paper uses Python language to mine, analyze and model the recruitment data in the Internet of Things database. The specific framework is shown in Figure 1.

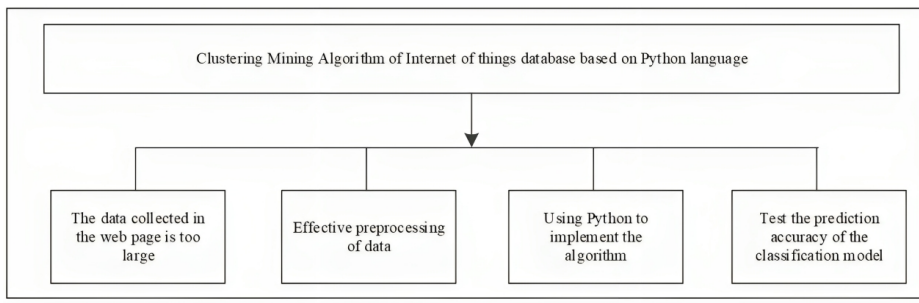


Figure 1. Algorithm framework

2.1 Improved Python-Based Scratch Crawler

In the traditional mode, the general open source crawler architecture of Sweep can not crawl the exclusive content. Therefore, the open source crawler structure based on Sweep is constructed, and the Bayesian classifier [10] is selected to judge the topic similarity. The clustering technique uses an automatic method in which the data is not labelled and also uses the solution to a challenge determined by the algorithm's expertise from a set of practice issues. Information grouping from all network members is the responsibility of every member node. Whenever a frame of information from every membership is obtained, the CH applies data acquisition before sending the frames towards the access point [11]. Through the search method, grab the Internet of Things database data in the Internet of Things database, analyze the recruitment data in the Internet of Things database through the clustering algorithm, maintain the node weights of the Internet of Things database at the present stage according to the analyzed contents, analyze the link weights of the child URLs according to the recruitment topic relevance module, and eliminate the topics less than the URLs node weights with the filter unit after calculation. Deduplicate the URL in the scratch structure, and determine the subject big data related to the Internet of Things database data through the following Bayesian classifier.

1. Set the set as $X = \{x_1, x_2, \ldots, x_m\}$, where the attribute in x is $a_1 - a_m$.
2. Set $Y = \{y_1, y_2, \ldots, y\}$ as the set of existing categories.
3. Calculate the probability $P(y_1 \mid x), P(y_2 \mid x), \ldots, P(y_n \mid x)$ in turn.
4. Calculate the maximum probability $P(y_k \mid x)$, the highest probability category has X and $x \in y_k$.

Through Bayesian theory, when $P(X \mid Y)$ is calculated, $P(Y \mid X)$ can be obtained. Therefore, Y is approximately equal to A and X is approximately equal to B. When B is divided into different attributes, each type of attribute is calculated in the following form:

1. Get a training sample set.
2. We can know the category of the sample set, and it is the set of items to be classified.
3. Through statistical analysis in the form of formula (1), the probability estimation value $P(x_m \mid y_n)$ of classification conditions is obtained.
4. If there are all characteristic attributes with independent conditions, it can be calculated by formula (1):

$$P(y_i \mid x) = \frac{P(y_k \mid x) P(y_i)}{P(x)}. \tag{1}$$

Since the denominator $P(x)$ is usually set as a constant, the maximized molecule can be calculated. At the same time, since each feature has conditionally independent attributes, therefore:

$$P'(y_i \mid x) = P(y_i) \prod_{k=1}^{n} P(y_k \mid x). \tag{2}$$

Through the above construction process of Bayesian classifier, we can judge the similarity of recruitment data topics in the Internet of Things database. When there are type feature attributes with larger weight, the higher the similarity of recruitment data topics in the Internet of Things database.

2.2 Data Preprocessing of Internet of Things Database

In this design, Python language will be used as the processing basis [12, 13], and this technology will be used to optimize the original data stream classification method. Individual categorization as well as ensemble categorization are the two techniques used to classify big quantities of information. The one-classification approach is quick and uses minimal storage during processing. At the same time, select the appropriate technology to complete the data flow processing, convert the original relational database into non-relational database, and improve the use effect of classification methods. Information is kept in a non-relational system, sometimes referred to as a NoSQL dataset. There are no columns, lines, cardinality, or database objects, except for database systems. Rather, models are appropriately tailored to the unique needs of the sort of data being saved used by the non-relational databases.

2.2.1 Text Data Stream Data Acquisition

The recruitment data in the Internet of Things database is complex and changeable, and the types vary greatly. When selecting target data, careful filtering and screening is required [14]. In this study, the data acquisition process is controlled by a formula, and the specific contents are as follows.

Set the unmarked data stream collected in the Internet of Things database, E contains N data records, and (x_w, y_w^{α}), where x_i represents a data label of w-dimensional attribute in the data, y_w^{α} represents the category label of w-dimensional attribute in the data, and α represents unknown items in the category. In order to build the recruitment data collection process in the Internet of Things database into the form of model, the data collection results are expressed as the set of marked data block F_i^g and unmarked data block F_i^h, and expressed by $E = \{F_1^1, F_2^g, \ldots, F_{m-1}^h, \ldots, F_m^n\}$.

Divide the collected i data blocks into positive data blocks and negative data blocks respectively. On the premise of ensuring the comprehensiveness of the data, set the first f data $f_1^1, f_2^g, \ldots, f_{m-1}^h, \ldots, f_m^n$ in the collected data as marked data

blocks [15, 16], and the remaining $f_{1+n}^{1}, f_{2+n}^{g}, \ldots, f_{m-n}^{h}, \ldots, f_{m}^{n}$ as unmarked data blocks. The types of data blocks can be divided by using formula (3):

$$y^{z} = P'\left(y_i \mid x\right) \arg\min\left\{\operatorname{dis}\left(x, y^{\varepsilon}\right), \operatorname{dis}\left(x, y^{n}\right)\right\}, \tag{3}$$

where $\operatorname{dis}\left(x, y^{2}\right)$ represents the distance between the unlabeled data block and the center point of the positive data block.

$$\operatorname{dis}\left(x, y^{8}\right) = \left\|x, y_1^{8}\right\|_2, \tag{4}$$

where, y_1^8 represents the center point of the positive data block. Similarly, the distance $\operatorname{dis}\left(x, y^{n}\right)$ between the unlabeled data block and the center point of the negative data block can be obtained.

According to the above process, the recruitment text data stream data in the Internet of Things database is obtained and used as the data basis in the subsequent classification process.

2.2.2 Data Stream Preprocessing and Mining

In order to better process the recruitment text data in the Internet of Things database, Python language and the Internet of Things database crawler technology [17] are applied to the process of data mining. Crawler search results include, for instance, Google and Yahoo. Crawlers have the benefit of containing a large quantity of documents. Search the keywords through the Internet of Things database crawler technology, get the corresponding data, integrate it into the form of database, and carry out preprocessing and mining. A web crawler, also referred as a retrieval system or internet spiders, is a software program utilized to systematically explore as well as analyze the contents of websites as well as similar online data. Because the Python language used in this design has a unique use environment, the original SQL database is transformed into PyMongo library, which is a non-relational database. In the Connections Graph, either right-click on a service, databases, or collecting and choose SQL Conversion from the context menu, or choose the SQL Conversion arrow in the taskbar. The suggested method for interacting with MongoDB from Python is PyMongo, a Python package that includes capabilities for doing so. Although MongoDB maintains information in BSON language both locally and remotely, users can still conceive of MongoDB as a JSON library. Different from the databases currently in use, PyMongo library can store text data from various sources and in various forms. Generally speaking, PyMongo offers a comprehensive selection of services that you are able to use for interaction with a MongoDB host. It offers the ability to access, extract data, update as well and remove information, in addition to executing network operations. At the same time, it improves the flexibility of the data model, makes the reading and writing of data easier, and facilitates the secondary expansion of data.

In this study, the text data is transformed into machine recognizable text and stored in vector space to complete the data stream preprocessing process. In order

to obtain the text features of the Internet of Things database data, the keywords in the data are extracted in the form of extraction according to the number and criticality of keywords [18, 19]. Data processing is the technological method of modifying information from another format, standard, or architecture to a different one without altering the information's contents. This is often done to make the information more usable by users or apps or to enhance the accuracy of the information. Access the URL string injected by SQL in the Internet of Things database, describe the string through a single text and divide the data in the form of word segmentation. After division, the space vector can be designed, which corresponds to the URL string. In this space, each feature can be described through the characteristic keywords added by SQL. The characteristics of SQL are simple to acquire. The following features make SQL a particularly useful and approachable vocabulary: wide range of instructions, database objects, movable vocabulary, reunites, unification, considered ideal for the client relationship, and interoperability. Among them, the inverse document rate can be determined by weight and analyzed according to word frequency [20].

1. **Word frequency:** Because the URL string is usually output in the form of ASCII code, and there is no rule in the string, when starting to segment the URL access data, it is necessary to segment the word in the form of "%", "&" and space, so that the data after the word segmentation can be left.

   If the keyword appears in each spatial vector feature, the number of times the keyword is found. The number of times after sorting is described by the word frequency. Keyword research is the method of locating these phrases. You will be capable of carrying out relevant keywords alone following completing this course. If there is no keyword, the word frequency is zero. The greater the number of keywords in the sample, the greater the importance. The process of adjusting the word frequency to the spatial vector model is represented by formulas (5) and (6):

$$T_1 = \frac{\text{tf}}{doc\_\text{length}}, \tag{5}$$

$$T_2 = \frac{tf}{\max ff}, \tag{6}$$

   where T_1 and T_2 represent features 1 and 2; tf indicates word frequency; doc_length indicates the length of the statement; max ff indicates the maximum word frequency in the statement. The process of transferring word frequency to spatial vector model can be carried out through formula (5) and formula (6).

2. **Reverse document rate:** Usually, the spatial vector is composed of word frequency and inverse document rate. Through the w-dimensional space, it can describe the spatial vector of each access data. The number of SQL injection keyword categories in the URL access data can determine the proportion S of

statements covering keywords in all statements. According to the above calculation of word frequency, the contribution degree of each word in the sentence can be obtained. When uploaded to the spatial vector, the interval degree between two sentences can be obtained [21]. When the number of keywords in a large number of data samples gradually decreases, it means that the data samples of the keyword are gradually reduced. Therefore, the higher the inverse document rate of the keyword, indicating that the word can distinguish sentences to a greater extent. The inverse document rate is described by I, therefore, I can be calculated by formula (7):

$$I = \ln\left(\frac{S}{df}\right) + 1. \tag{7}$$

If you want to realize that the reverse document rate is within the $[0, 1]$ range, you can calculate it through formula (8):

$$I_{new} = \frac{\ln\left(\frac{S+0.5}{df}\right)}{\ln S}. \tag{8}$$

Through formula (8), the calculation of inverse document rate can be completed, and its value represents the feature vector of a keyword.

Since the data in the form of a local domain name is saved, the data in the form of HTML source code will be cleaned for the data in the form of local domain name. According to the text space vector model constructed in this paper, the word frequency features of Internet of Things database data are extracted, and the final form is stored in HDFS (Hadoop Distributed File System) distributed database. An algebraic paradigm called the vector space paradigm treats items (including words) as coordinates. This enables determining word resemblance or the relevancy of a user's query or content simply.

In the process of data flow cleaning, deal with the abnormal items in the data collection, correct the data whose attribute value is missing and is not related to the data flow due to other reasons and correct the errors in the data collection. Because the research object is the Internet of Things database data, its types are more complex. Therefore, outliers in the data stream need to be processed to ensure the accuracy of data classification. In data cleaning, the data with large differences in average value in the data stream shall be eliminated to avoid data pollution. The most common causes of poor data integrity are improper data entry typos, erroneous material, misleading data, and so on. Each organization needs to perform data analysis. At the same time, set the corresponding data rules to restrict the data preprocessing results and data mining results. For multidimensional data such as recruitment data in the Internet of Things database, data preprocessing will relatively reduce the data dimension, resulting in the inability to accurately define the data in the process of data analysis. When this data reaches a certain amount, the high latitude of the data

will increase the difficulty of data analysis and make the overall proportion of the data unbalanced. Organizations have the difficulty of expanding analysis process due to the continually growing storage capacity. As the information accumulates, it gets more and more complex to analyze and provide various documents. Therefore, in the process of data mining, adaptive classification method is selected to improve the accuracy of clustering mining of the Internet of Things database.

3 ADAPTIVE CLASSIFICATION OF TEXT DATA STREAM

The traditional supervised learning algorithm has poor data ability for data, which is easy to cause a huge waste of data resources. Using machine learning algorithm to complete the classification of data flow, the classifier has low generalization ability, which can effectively improve the effectiveness of data flow processing. The method of semi supervised learning and semi cluster analysis [22, 23] is used to construct the corresponding classifier. Semi-supervised aggregation techniques are those that can be used with partly labelled information or data with different kinds of outcomes metrics. Creating groupings or clustered whilst making sure that the data are as comparable as feasible in each category is the overall goal of hierarchical clustering in advertising. Semi supervised learning is a new method in the field of machine learning. For a given data preprocessing result get J, it can be expressed as:

$$J = (x_1, y_2), \ldots, \left(x_{|g|}, y_{|g|}\right), \left(x_{|g|+1}, y_{|g|+1}\right), \ldots, \left(x_{|g|+|h|}, y_{|g|+1}\right), \tag{9}$$

where, $|g|$ represents the number of labeled samples, $|h|$ represents the number of unmarked samples, and $|h| + |g| = n$. Semi supervised learning is to classify and process the data with missing identification in the sample. By service-learning, individuals acquire to think critically as well as consider their experiences. These abilities include the capacity to connect seemingly unrelated aspects of an encounter in significant manner, to look for trends as well as higher purpose in data, or to assess and evaluate situations. The data is shown in Figure 2.

According to the above formula, assuming that the marked data and ummarked data in the data are mixed distributed according to Gaussian distribution, then:

$$r(x \mid \delta) = \delta \sum_{i=1}^{n} \varepsilon_i \left(x_{|g|+|h|}, y_{|g|+1}\right), \tag{10}$$

where ε_i represents the mixing coefficient of the data stream and δ represents the distribution coefficient. The Bayes Theorem is the foundation of Bayesian categorization. The economic analyzers are Bayesian predictors. The likelihood that a specific packet corresponds to a specific class is one example of a class participation likelihood that can be predicted by Bayesian processors. According to Bayesian classification theory, if the recruitment data is class q in the data flow, the posterior

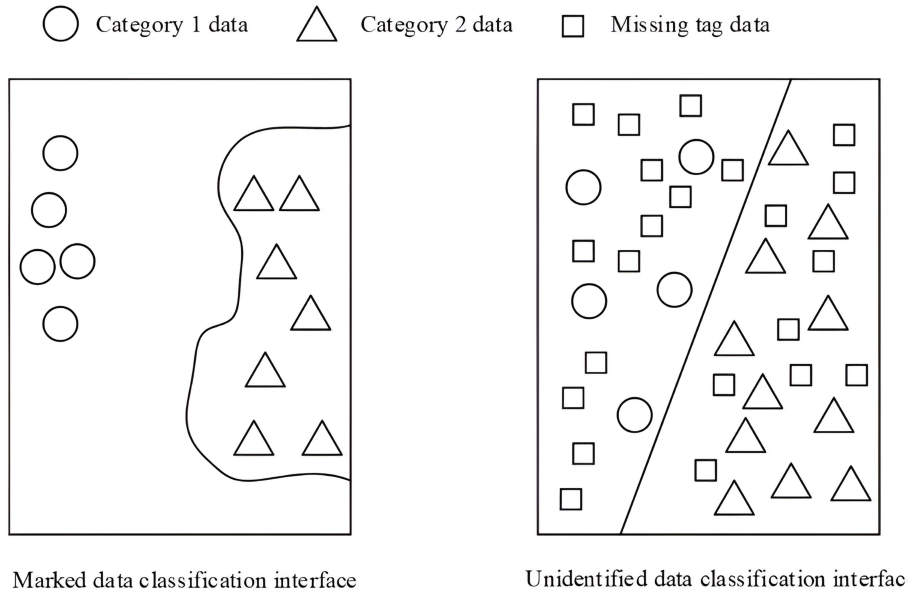


Figure 2. Data flow classification hyperplane

probability is the largest. The specific calculation formula is formula (11):

$$p = \frac{\varepsilon_i r(x \mid \delta)}{y^x \sum_{i=1}^{n} r(x \mid \delta)} \tag{11}$$

The unmarked data in the data stream is estimated through formula (11), and the corresponding ummarked data is obtained through the marked data, so as to improve the accuracy and ensure the use effect of the classifier.

A dynamical structure is found to suit a particular message or generalized linear of information is a time series description, often referred to as a signaling prototype. Multimodal algorithms can be created when the time series is multidimensional. In addition to the above process, according to the big data text features extracted by the text space vector model, a time series model is designed to arrange the data to remove the delay in the data. Let Z_i be the multidimensional random variable at this stage, and $Z = \{Z_1, Z_1, \ldots, Z_N\}$ represents the number of recruitment data in the Internet of Things database at this stage. When it is within the data value range of routing link layer, calculate the similarity characteristics of recruitment data through formula (12):

$$\rho = \varpi\left(x, y^g\right) + p \prod_{k=1}^{n} u\mathrm{P}\left(y_k \mid x\right), \tag{12}$$

where u and ϖ represent the response frequency and baud rate of the characteristic sequence exclusive to the recruitment data. According to formula (12), the time series can be constructed through time nodes, so as to eliminate the delay characteristic. Differentiating is arguably the easiest way to detrend a linear model. The quantity at the present rate phase is determined as the differential among the initial remark as well as the measurement at the prior data increment, namely, and a novel sequence is created.

The above contents are integrated and integrated with the adaptive classification method of data flow. So far, the design of clustering mining algorithm of the Internet of Things database based on Python language is completed.

4 EXPERIMENTAL ANALYSIS

4.1 Experimental Setup

This method is applied to a recruitment website to crawl the recruitment big data of the website, and the performance of this method is verified by experiments. At the same time, the data mining clustering method based on multiple minimum support of pattern growth in reference [3] and the data mining clustering method based on response threshold search in reference [5] are compared with this method. Minsup, or sequential pattern extraction, operates under the implied premise that each object has the identical quality (i.e., frequency). The regulations concerning rare Things will not be recognized if the minsup level is sufficient. Reaction criteria describe the probability of responding to stimulus related to a given task. People with low thresholds do activities with less stimuli than people with high thresholds.

In order to test the clustering mining effect of this method, the experimental test is carried out by using Windows 10 operating system. The specific experimental environment is as follows: Hardware environment: PC: CPU 30 GHZ, 8 GB memory and 500 GB hard disk; Development language: using Java language, it is relatively simple, flexible, portable and superior in performance.

Development tools: Eclipse is an extensible development platform with the advantages of easy operation. It is a development tool often used by Java.

Database development tool: MS SQL Server 2008, with high intelligence and high development efficiency. In the above experimental environment, the proposed algorithm is used to crawl the data from the recruitment website and take it as the data source of this experiment. The URLs on the website that a web search crawler may visit are specified in a robots.txt file. This is mostly intended to prevent the website from becoming overloaded with queries; it is not a method of taking a website off from Search. First, the crawler sends an access request and automatically saves its content if it meets the access requirements; Secondly, the crawler analysis module is used to obtain the remaining links in the crawled pages and take them as the subsequent crawling targets. Web crawlers are applications created to traverse the network, obtain information, organize it, then analyze it to speed up results. Crawlers are a clever answer to huge amounts of information and a driving

force behind significant developments in the cyber security sector. The web crawler framework is shown in Figure 3.

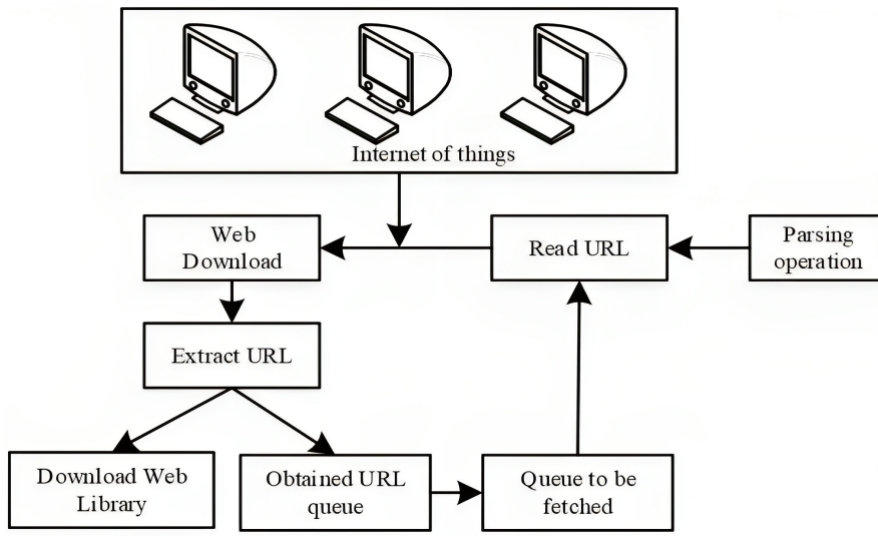


Figure 3. Schematic diagram of crawler frame

Similar to the working principle of users using browsers, the crawler crawling process also processes requests. Search engine crawlers use the technique of crawling to find provided by web resources. Whenever a person enters a search function, the appropriate outcomes are displayed on the web browser due to crawling, which is the process by which search engine crawlers scan the web sites or store a record of every data on indexed databases. Take the operation of rendering a web page by the browser as an example. When a user opens a recruitment web page, the browser will initiate a request, and the server will respond to the request. At the same time, the browser will display the parsing request on the page.

Develop crawler workflow through crawler framework. A web crawler finds URLs, then reads, analyses, and categorizes web sites. They discover connections to other websites anywhere along route as well as add those to the database of websites to crawl later. Select the recruitment web page with high value as the initial running target, and save the web page address and its corresponding path name to the downloader. For the recruitment webpage downloaded to the local, it is processed in three parts: first, save it to the preset page library and wait for index processing; Second, mark the page as a crawling page and add it to the crawling queue; Third, analyze the downloaded recruitment page and compare it with the capture queue. If there is no such page in the queue, transfer the recruitment page out of the queue and continue to wait for the scheduling instruction. According to the above process, the crawler runs automatically, completes all crawling work, and obtains

the recruitment data mining results. The whole process of clustering and mining recruitment data in the Internet of Things database using the above procedures is shown in Figure 4.

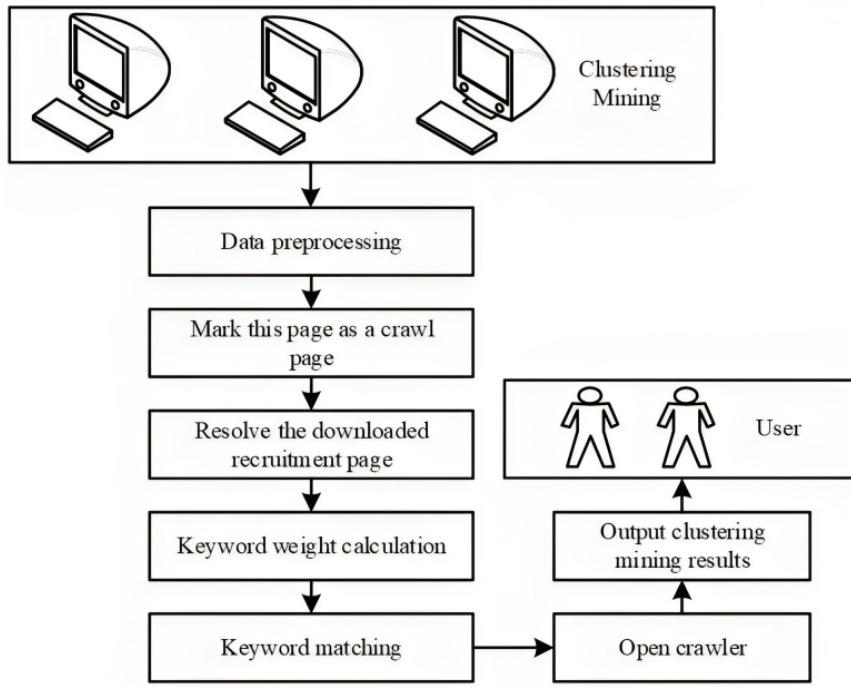


Figure 4. Schematic diagram of recruitment data clustering mining in the Internet of Things database

According to the process shown in Figure 4, obtain recruitment data and set that each data flow contains multiple and different numbers of sample data blocks. The specific data information is shown in Table 1.

	Test Data Set 1	Test Data Set 2	Test Data Set 3
Number of data blocks/piece	30	60	50
Number of attributes/piece	20	25	35
Number of concepts/piece	6	7	5
Number of time points/piece	45	25	30
Total data/piece	650	750	1 150
Important information/article	15	25	30

Table 1. Test data set setting

4.2 Result Analysis

Analyze the crawling ability of recruitment data of the Internet of Things database with different methods, and conduct experiments for various indicators. It all comes down to gathering and evaluating data utilizing modelling, computer vision, and statistical methods in order to provide the most accurate predictions about what might occur in certain situations. The experimental results are shown in Table 2.

Data Crawling Capacity	Paper Method	Reference [3] Method	Reference [5] Method
Climbing speed	About 300 pieces/min	About 270 pieces/min	About 260 pieces/min
Pertinence	Very high, able to crawl specific data	High, more data can be obtained	High, easy to lose some important data
Expansibility	Strong ability to crawl data associated with keywords	Strong, only the specified keywords can be crawled	Strong
Generality	Very strong	Strong	Strong for simple changes
Flexibility	Very strong	Strong	Strong, can only change with the changes of the web page
Crawling performance	It has strong crawling performance and can parse html files and xml files at the same time	When parsing html pages, you need to use other libraries	It mainly constructs the web scraper, and the crawling performance is poor
Crawling ability	Frame structure, which can crawl to the corresponding web page data and structured data, and the crawl speed is fast	It can extract data, but it cannot be used as a crawler independently	The speed of crawling the corresponding web page data is slow

Table 2. Recruitment data crawling capacity of the Internet of Things database with different methods

According to Table 2, through the analysis of data crawling ability of different methods, the data crawling speed of this method is fast, and about 300 data can be crawled per minute, while the crawling quantity of the other two methods is lower than that of this method in every minute, and the crawling performance of this method is higher than that of other methods. Therefore, this method can effectively realize the crawling of recruitment data in the Internet of Things database, provide a favorable basis for the subsequent mining of this kind of big data.

Basement, open-surface (pit), subsurface, as well as in-situ mining are the four basic types of extraction. Since it takes a certain time to mine, five feature data groups are selected to analyze the reading delay of different mining methods through the comparison of three methods. The experimental results are shown in Figure 5.

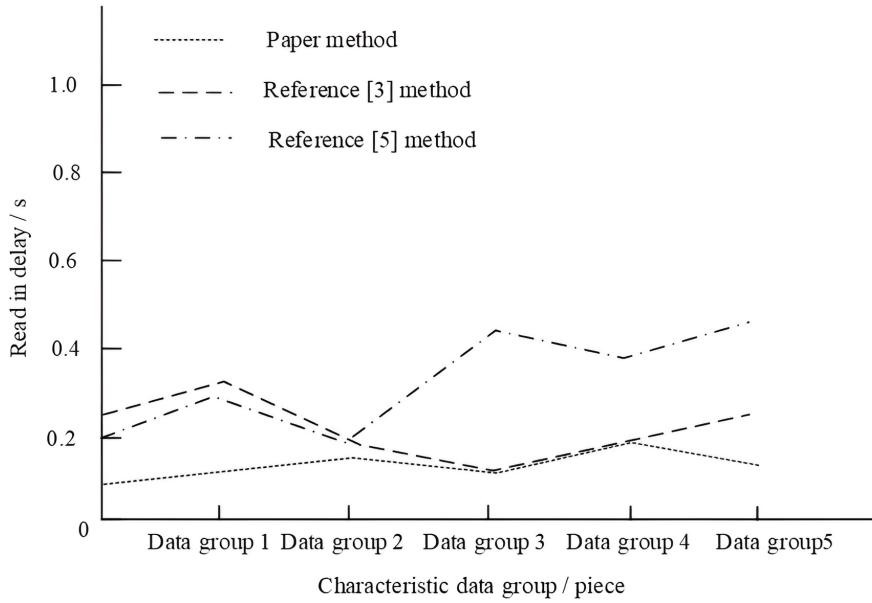


Figure 5. Comparison of read in delay of different methods

It can be seen from Figure 5 that due to the inconsistency of data in each group, there are different reading delays in each group. Among them, the reading delay of reference [5] method in data group 3, data group 4 and data group 5 remains the highest, up to 0.8 s, while the reading delay of reference [3] method in data group 1 and data group 2 is the highest, up to more than 0.4 s, while the method in this paper maintains the lowest reading delay in each group of data mining, And they are kept below 0.2 s. Therefore, this method can quickly realize big data mining.

An experiment is conducted on the data keywords of the Internet of Things database mined by using this method. By analyzing the actual attention of the keyword in different dates and the attention obtained by using this method, the big data mining ability of this paper is obtained. The analysis results are shown in Figure 6.

According to Figure 6, with the gradual increase of the date, the attention of the recruitment data gradually increases. When it reaches the 15^{th} day, the attention of the recruitment data gradually decreases until the recruitment data is no longer

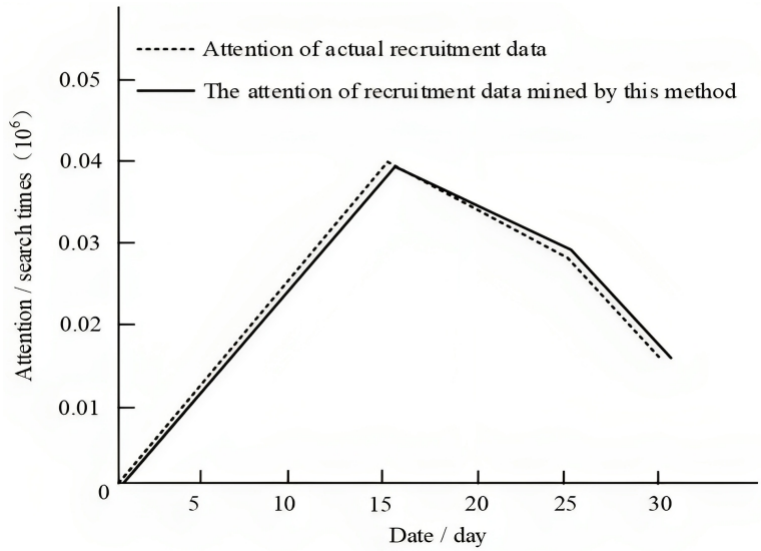


Figure 6. Analysis of recruitment data attention

concerned. The test shows that the attention of the recruitment data mined by this method is almost in line with the actual attention, and there is no large deviation. Therefore, this method can effectively mine the recruitment data in the Internet of Things database.

According to the number of posts generated by recruitment data keywords in the Internet of Things database in 24 hours, analyze the number obtained after mining with different methods. The analysis results are shown in Figure 7.

According to Figure 7, the number of new posts in a day changes from increase to decrease with time, and the number of new posts is the highest around 11:00 and 19:00. After mining by different methods, the change of the number of posts obtained by this method is completely similar to that of the actual posts, while the number of posts obtained by the method of reference [5] does not meet the standard of the actual number of posts, The number of posts obtained by the method of reference [3] is completely higher than the actual number. Therefore, the mining process of this method is more accurate and can fully mine the number of new posts associated with this keyword.

Analyze the click time frequency within 1 h after the recruitment keyword is generated, and compare it with the click situation obtained after mining by this method. The analysis results are shown in Figure 8.

According to Figure 8, the click time frequency of the keyword gradually increases within 1 h and decreases after 30 min. The mining results of this method are consistent with the actual click time frequency, which can clearly mine the click situation of the keyword.

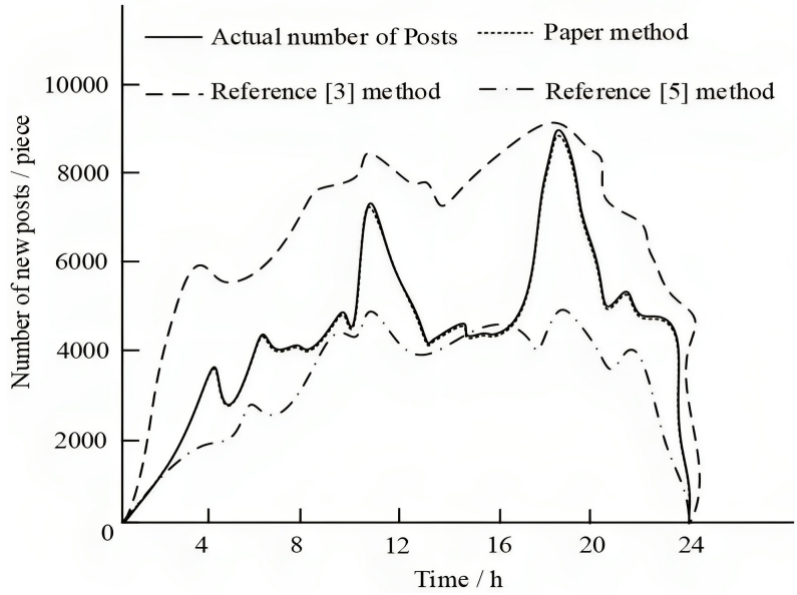


Figure 7. Number of new posts mined by different methods

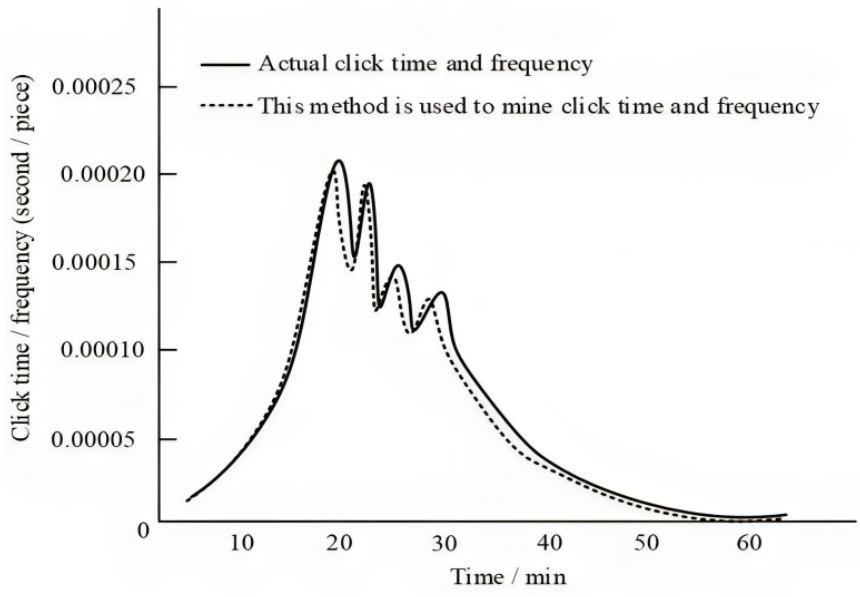


Figure 8. Comparison of keyword click time and frequency

Five retrieval experiments were conducted in the recruitment website to obtain the accuracy and recall of the three methods. Recall is the modeling measure we utilize to choose the optimal platform whenever there is a massive expense connected with False Negative, according to the identical concept. Several items and components can be structured using packaging, that contributes to a well-organized structure of information and makes subdirectories or sections simple to retrieve. The experimental results are shown in Table 3.

Number of Experiments	Paper Method		Reference [3] Method		Reference [5] Method	
	Accuracy [%]	Recall Rate [%]	Accuracy [%]	Recall Rate [%]	Accuracy [%]	Recall Rate [%]
1	90.6	63.1	83.2	60.1	82.7	61.4
2	94.8	68.2	87.3	63.9	80.3	66.3
3	93.1	72.8	78.8	62.2	82.9	69.9
4	92.2	72.7	79.6	60.2	89.2	65.3
5	95.9	78.8	80.9	65.9	81.5	68.1

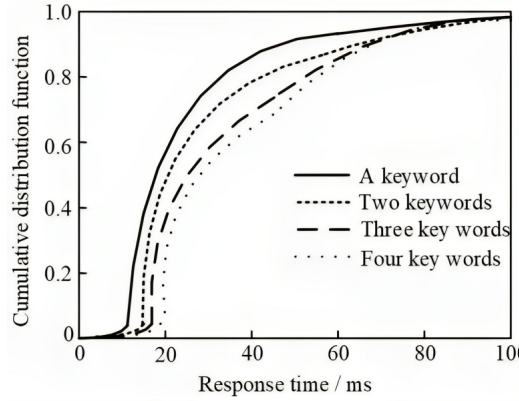
Table 3. Retrieval accuracy and recall of different methods

It can be seen from Table 3 that in the five retrieval tests, the performance of this method is more stable, and the accuracy and recall are significantly higher than the other two methods. This is because before the start of the crawler, all resource tags are preprocessed to remove low-frequency words and stop words, retain key words, and improve the retrieval accuracy to a certain extent. When the above experimental conditions remain unchanged, the response time of clustering mining of the three methods is compared, and the results are shown in Figure 9, respectively.

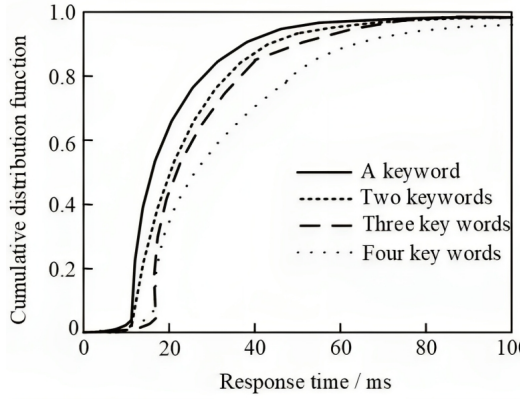
As can be seen from Figure 9 (9 a), 9 b) and 9 c)), the cumulative distribution function curves of the three methods show a smooth beginning, then rise significantly, and finally reach the mean value. However, when there are fewer keywords in the methods of reference [3] and reference [5], the required response time is shorter, but the clustering mining delay increases with the increase of keywords. The method in this paper is not sensitive to the number of keywords. It will get a response at about 40 ms, and there is almost no significant difference between the four curves. This is because the crawler route design is more reasonable, the crawler route with low weight is removed, and the response speed of clustering mining is accelerated.

5 CONCLUSION

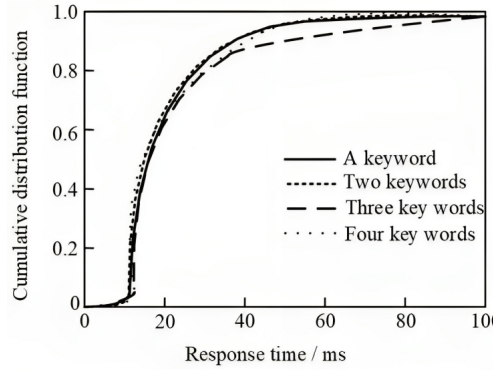
In this paper, by using Python language tools, combined with the relevant resource content existing in Python language, the improved Python based scratch crawler crawls the recruitment data existing in the Internet of Things database, extracts the recruitment data keywords through the text space vector model, mines the recruitment data through the time series model that eliminates the characteristic



a) Reference [3] method cluster mining response time



b) Reference [5] method cluster mining response time



c) Paper method cluster mining response time

Figure 9. Comparison curves of numerical solutions at different times

delay, and finally realizes clustering mining through the clustering algorithm. Finally, taking the data mining of recruitment website as an experimental case, the effectiveness of this method is verified. This method can continue to be optimized in the future, so that the Internet of Things database data can be positioned and analyzed more accurately.

6 DECLARATIONS

Funding: Science and Technology Research Project of Jiangxi Provincial Department of Education: Research on Clustering Mining Algorithm of Internet of Things Database Based on Python Language (Project No. GJJ218415).

Conflict of Interest: There is no conflict of interest among the authors.

Data Availability: All data generated or analyzed during this study are included in the manuscript.

Code Availability: Not applicable.

Author's Contributions: Fang Wan and Ying Liu contributed to the design and methodology of this study, the assessment of the outcomes and the writing of the manuscript.

REFERENCES

[1] Xi, J.—Tang, J. L.: Research on Task Slot Scheduling of Internet of Things Terminal Based on Dynamic Topology. Computer Simulation, Vol. 39, 2022, No. 1, pp. 366–369 (in Chinese).

[2] Vandebon, J.—Coutinho, J. G. F.—Luk, W.—Nurvitadhi, E.: Enhancing High-Level Synthesis Using a Meta-Programming Approach. IEEE Transactions on Computers, Vol. 70, 2021, No. 12, pp. 2043–2055, doi: 10.1109/TC.2021.3096429.

[3] Zhu, A.: Spatiotemporal Feature Mining Algorithm Based on Multiple Minimum Supports of Pattern Growth in Internet of Things. The Journal of Supercomputing, Vol. 76, 2020, No. 12, pp. 9755–9771, doi: 10.1007/s11227-020-03217-x.

[4] Amini, S. M.—Karimi, A.: Two-Level Distributed Clustering Routing Algorithm Based on Unequal Clusters for Large-Scale Internet of Things Networks. The Journal of Supercomputing, Vol. 76, 2020, pp. 2158–2190, doi: 10.1007/s11227-019-03067-2.

[5] Zhou, Q.—Hao, J. K.—Wu, Q.: Responsive Threshold Search Based Memetic Algorithm for Balanced Minimum Sum-of-Squares Clustering. Information Sciences, Vol. 569, 2021, pp. 184–204, doi: 10.1016/j.ins.2021.04.014.

[6] Castellanos-García, L. J.—Elci, S. G.—Vachet, R. W.: Reconstruction, Analysis, and Segmentation of LA-ICP-MS Imaging Data Using Python for the Identification of Sub-Organ Regions in Tissues. Analyst, Vol. 145, 2020, No. 10, pp. 3705–3712, doi: 10.1039/C9AN02472G.

[7] Kitson, E.—Kew, W.—Ding, W.—Bell, N. G. A.: PyKrev: A Python Library for the Analysis of Complex Mixture FT-MS Data. Journal of the American Society for Mass Spectrometry, Vol. 32, 2021, No. 5, pp. 1263–1267, doi: 10.1021/jasms.1c00064.

[8] Schmartz, G. P.—Hartung, A.—Hirsch, P.—Kern, F.—Fehlmann, T.—Müller, R.—Keller, A.: PLSDB: Advancing a Comprehensive Database of Bacterial Plasmids. Nucleic Acids Research, Vol. 50, 2022, No. D1, pp. D273–D278, doi: 10.1093/nar/gkab1111.

[9] Karpachevskiy, A.—Titov, G.—Filippova, O.: Development of a Spatiotemporal Database for Evolution Analysis of the Moscow Backbone Power Grid. Data, Vol. 6, 2021, No. 12, Art. No. 127, doi: 10.3390/data6120127.

[10] Halbersberg, D.—Wienreb, M.—Lerner, B.: Joint Maximization of Accuracy and Information for Learning the Structure of a Bayesian Network Classifier. Machine Learning, Vol. 109, 2020, No. 5, pp. 1039–1099, doi: 10.1007/s10994-020-05869-5.

[11] Manogaran, G.—Srivastava, G.—Muthu, B. A.—Baskar, S.—Shakeel, P. M.—Hsu, C. H.—Bashir, A. K.—Kumar, P. M.: A Response-Aware Traffic Offloading Scheme Using Regression Machine Learning for User-Centric Large-Scale Internet of Things. IEEE Internet of Things Journal, Vol. 8, 2020, No. 5, pp. 3360–3368, doi: 10.1109/JIOT.2020.3022322.

[12] Popko, V. V.—Sentemova, D. V.—Yarmalovich, M. A.: Optimized Reproduction, Collection, Transmission, Processing, and Display of Measurement Data Obtained Using the National Standard of Voltage Unit – Volt. Measurement Techniques, Vol. 64, 2021, No. 7, pp. 556–561, doi: 10.1007/s11018-021-01978-2.

[13] Azarafza, M.—Koçkar, M. K.—Faramarzi, L.: Spacing and Block Volume Estimation in Discontinuous Rock Masses Using Image Processing Technique: A Case Study. Environmental Earth Sciences, Vol. 80, 2021, No. 14, Art. No. 471, doi: 10.1007/s12665-021-09768-3.

[14] Aslam, A.—Chen, H.—Jin, H.: Pre-Filtering Based Summarization for Data Partitioning in Distributed Stream Processing. Concurrency and Computation: Practice and Experience, Vol. 33, 2021, No. 20, Art. No. e6338, doi: 10.1002/cpe.6338.

[15] Liu, S.—Ding, F.—Yang, E.: Iterative State and Parameter Estimation Algorithms for Bilinear State-Space Systems by Using the Block Matrix Inversion and the Hierarchical Principle. Nonlinear Dynamics, Vol. 106, 2021, No. 3, pp. 2183–2202, doi: 10.1007/s11071-021-06914-1.

[16] Burns, S.—Collisson, E. A.: Blockchain-Authenticated Sharing of Cancer Patient Genomic and Clinical Outcomes Data. Journal of Clinical Oncology, Vol. 38, 2020, Art. No. e19358, doi: 10.1200/JCO.2020.38.15_suppl.e19358.

[17] Iggena, T.—Bin Ilyas, E.—Fischer, M.—Tönjes, R.—Elsaleh, T.—Rezvani, R.—Pourshahrokhi, N.—Bischof, S.—Fernbach, A.—Parreira, J. X. et al.: IoTCrawler: Challenges and Solutions for Searching the Internet of Things. Sensors, Vol. 21, 2021, No. 5, Art. No. 1559, doi: 10.3390/s21051559.

[18] Kim, K. O.—Jun, B. J.—Lee, B.—Park, S. J.—Roh, G.: Comparison of First Criticality Prediction and Experiment of the Jordan Research and Training Reactor

(JRTR). Nuclear Engineering and Technology, Vol. 52, 2020, No. 1, pp. 14–18, doi: 10.1016/j.net.2019.06.027.

[19] Kislyi, A. A.—Ravkin, Y. S.—Bogomolova, I. N.—Tsybulin, S. M.—Starikov, V. P.: Number and Distribution of the Narrow-Headed Vole Lasiopodomys gregalis (Pallas, 1779) (Cricetidae, Rodentia) in Western Siberia. Biology Bulletin, Vol. 48, 2021, No. 10, pp. 1822–1831, doi: 10.1134/S1062359021100162.

[20] Rivero-Contreras, M.—Engelhardt, P. E.—Saldaña, D.: An Experimental Eye-Tracking Study of Text Adaptation for Readers with Dyslexia: Effects of Visual Support and Word Frequency. Annals of Dyslexia, Vol. 71, 2021, pp. 170–187, doi: 10.1007/s11881-021-00217-1.

[21] Imran, S. S.—Zainab, H.: Degree of Separation Between Articular Facets (Anterior and Middle) on Anterior Third of Superior Surface of the Calcaneum in H.K.E Region. International Journal of Anatomy and Research, Vol. 8, 2020, No. 3.2, pp. 7633–7638, doi: 10.16965/ijar.2020.124.

[22] Wei, J.—Du, W.—Zhang, B.—Cui, J.—Yi, J.—Kang, S.—Li, J.—Feng, C.: Effect of Interval Cooling Time Between Weld Passes on Temperature Field of Inconel 625. Ordnance Material Science and Engineering, Vol. 43, 2020, No. 6, pp. 90–94 (in Chinese).

[23] Sahoo, P.—Roy, I.—Ahlawat, R.—Irtiza, S.—Khan, L.: Potential Diagnosis of COVID-19 from Chest X-Ray and CT Findings Using Semi-Supervised Learning. Physical and Engineering Sciences in Medicine, Vol. 45, 2022, No. 1, pp. 31–42, doi: 10.1007/s13246-021-01075-2.

Fang Wan is a Lecturer. She had received her Bachelor Degree in 2004 from the Nanchang University in the Department of Software Engineering. Then, she was awarded the Master degree in 2008 from the Nanchang University in the Department of Software Engineering. Now she is working in the Nanchang JiaoTong Institute. Her research interests include electronic commerce and data analysis. She has published one academic paper. Meanwhile, she is also responsible for compiling several textbooks. She also participated in two cooperative research projects.

Ying Liu is an Associate Professor. She received her Bachelor Degree in 2004 from the Nanchang University in the Department of Software Engineering. Now she is working in the Nanchang JiaoTong Institute. Her research areas include database and computer graphics. She has published two academic papers. Meanwhile, she also participated in science and technology project of the Jiangxi Provincial Department of Education.