Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 4, pp. 2636–2644, DOI 10.12694/scpe.v25i4.2927

HETEROGENEOUS HIGH PERFORMANCE DATA MINING SYSTEM FOR INTELLIGENT DATA

XINKE WANG*, KAI LI†, AND XIAOLING LI‡

Abstract. In order to improve the utilization rate of internet data under heterogeneous distribution, increase the diversified usage functions and data transmission rate of the internet, and reduce the running time of the internet, it is necessary to mine internet data under heterogeneous distribution. The author proposes an ontology based optimization method for internet data mining under heterogeneous distribution; This method first preprocesses and selects data features from internet data under heterogeneous distribution, and uses a feature selection decision system to select features from the mining data. Based on this, information entropy is used to filter internet data under heterogeneous distribution. During the filtering process, the theoretical values filtered by information entropy are reduced to obtain the optimal data filtering value, finally, based on the various data information obtained in the preprocessing, the iterative calculation results of the information gain value in the decision tree generation algorithm are used to high-precision mine internet data under heterogeneous distribution; The simulation experimental results demonstrate that the proposed method improves the flexibility of internet data operations under heterogeneous distribution, increases the recyclability of internet data, and makes internet operations under heterogeneous distribution more concise and efficient, providing a strong basis for research and development in this field.

Key words: Heterogeneous distribution; Internet; Data mining methods; Optimization research

1. Introduction. Since the 1990s, "informatization", "networking", and "globalization" have become several main characteristics of human socio-economic development in the new era [1]. Information and data have become important resources that support human socio-economic development after material and energy, and are also changing the allocation of social resources, as well as human values, work, and lifestyle. With the rapid development of information technology, fundamental changes have taken place in various industries driven by information technology. Various information systems are abundant, and governments, enterprises, and research institutions have established a large number of information systems.

Data is growing at an annual rate of 200%, with most of it coming from the application of new technologies. Due to differences in business and functions, as well as the impact of factors such as stage, technical, and human factors on the informationization construction of various departments and institutions, the established information systems are isolated from each other, forming multiple "information islands". These "information silos" make the internal data of enterprises exhibit obvious characteristics such as distribution, self-control, and heterogeneity. With the development of information technology, information and data have become important assets in today's society and a major driving force for its development. It is in this situation that research on information and data is also increasing, and the breadth and depth of research are further deepened. Enterprises are not satisfied that the system can only provide business data for local business processes, but increasingly need to achieve information sharing among multiple businesses distributed in different locations on the network to improve the efficiency of enterprise collaborative operation. Therefore, in order to ensure the sharing, maintenance, and management of information within and between enterprises, it is necessary to find a unified operation method for multi-source and heterogeneous data. Heterogeneous data integration is a key problem that must be solved in information integration applications.

Heterogeneous data integration has important characteristics, mainly manifested as not only shielding the distribution and heterogeneity of heterogeneous data, but also fully maintaining the autonomy of heterogeneous

*Zhengzhou Technical College, Zhengzhou, 450121, China
†Zhengzhou Technical College, Zhengzhou, 450121, China
‡Zhengzhou University of Economics and Business, Zhengzhou, 450000, China (Corresponding author, lxl2004lxl22@163.com)

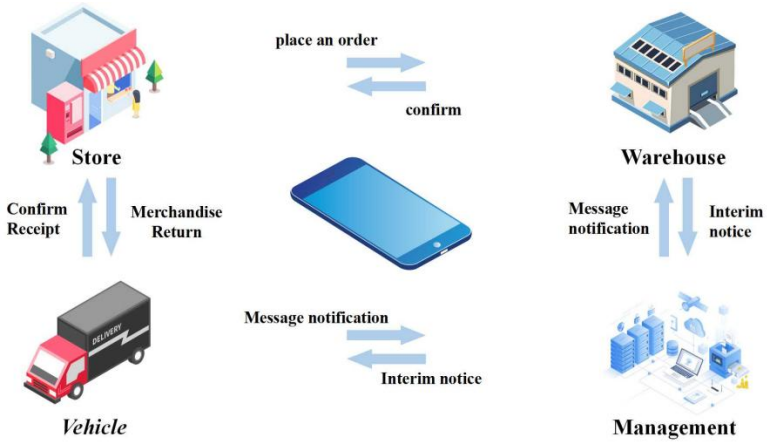

Fig. 1.1: Data integration method for multi-source heterogeneous massive data

data. The ultimate result is that users do not have to worry about the physical storage location of heterogeneous data, only need to operate and use heterogeneous data through an integrated environment, and do not need to worry about the structural differences between data. At the same time, the integrated system does not affect various local application systems. Heterogeneous data integration systems provide a solution for enterprises to integrate multiple platforms, applications, structures, and semantic data. Through such an integrated system, not only can various relevant data resources within the enterprise be integrated, but also external information can be collected for data mining to provide support for enterprise decision-making. Therefore, heterogeneous data integration systems are receiving increasing attention, and research in this area has become a hot topic. Figure 1 shows a data integration method for multi-source heterogeneous massive data.

With the continuous development and popularization of computer science and internet technology, internet data under heterogeneous distribution is distributed in office automation, electronic data exchange, remote exchange, remote education, electronic bulletin board system BBS, electronic banking, securities and futures trading, broadcasting packet exchange, information superhighway, enterprise network, etc Intelligent buildings and structured cabling systems are widely used in social platforms and systems [2]. Therefore, the development of internet data under heterogeneous distribution has received widespread attention and high attention from people. The internet under heterogeneous distribution can not only support different applications of different protocols and combine advantageous products or systems, but also meet the diverse needs of network business and improve the utilization rate of various multi-functional platforms and systems in the internet. Due to the characteristics of uncertainty, diversity, and flexibility of the internet under heterogeneous distribution, it is necessary to mine internet data under heterogeneous distribution [3,4]. Most internet data mining methods under heterogeneous distribution are unable to quickly, effectively, and accurately mine data, resulting in high packet loss rates, complex data operation processes, and computational errors during operation or operation of the internet under heterogeneous distribution. In this situation, how to reduce the packet loss rate of internet data mining under heterogeneous distribution and improve the accuracy of internet data mining has become an urgent problem to be solved. The optimization method of in ternet data mining based on ontology under heterogeneous distribution can perform flexible, convenient, reliable, and high-precision data mining on it. It is a feasible way to solve the above problems [5]. In response to the aforementioned issues, the author proposes an ontology based optimization method for internet data mining under heterogeneous distribution. This method first preprocesses the internet data under heterogeneous distribution to achieve higher mining accuracy and faster mining speed. Then, the decision tree generation algorithm is used to mine the internet data under heterogeneous distribution. Simulation experiments have shown that the proposed method can efficiently and accurately mine internet data under heterogeneous distribution, and has good implementability.

2. Optimization Method for Internet Data Mining under Heterogeneous Distribution.

2.1. Internet Data Preprocessing under Heterogeneous Distribution. Using ontology to mine internet data under heterogeneous distribution, the first step is to preprocess internet data under heterogeneous distribution. In data preprocessing, it is necessary to determine the data target attribute set and data condition attribute set of the internet original dataset under heterogeneous distribution, secondly, the value range of the attribute set is divided into several cells, and a discrete symbol of internet data corresponds to a data attribute set between cells. This results in a feature selection decision system for internet data under heterogeneous distribution. The same data records in the feature selection decision system are merged and recorded as (R, CRD) to establish a feature selection decision system for internet data under heterogeneous distribution, it is an optimization of the traditional data mining methods that directly conduct internet data mining without data feature selection [6,7].

Before data feature selection, the features of the data need to be extracted, and the author uses the maximum interval algorithm to extract features from internet data under heterogeneous distribution. Assuming that internet data follows a certain feature distribution P, the similarity value ω of internet data feature extraction under heterogeneous distribution is calculated using the maximum interval algorithm as follows:

$$\omega = arg_{\omega}\omega_q^0 \tag{2.1}$$

Among them, 0 represents the predefined threshold for feature extraction from internet data under heterogeneous distribution, and q represents the dimensionality of internet data feature extraction. Based on the extraction of similarity value ω from internet data features under heterogeneous distribution, it is known that the extraction process of internet data features is as follows:

Input F,G,h.
Output: $\omega_1, \omega_2, \omega_3$.

Among them, F represents the internet data feature extraction dataset under heterogeneous distribution, G represents a feature parameter in the internet data feature extraction under heterogeneous distribution, and h represents the dimensionality of the features to be extracted from the internet data under heterogeneous distribution, $\overline{o}$ represents the feature attribute mapping values in internet data feature extraction [8]. This completes the feature extraction of internet data under heterogeneous distribution. Applying ontology to feature selection of internet data under heterogeneous distribution aims to extract the most essential features that reflect the essence of internet data mining under heterogeneous distribution from the original internet data under heterogeneous distribution. The following is the specific process of feature selection methods for internet data under heterogeneous distribution:

Assuming input: internet data condition attribute set A, data decision attribute set B, and data decision system (R, CRD) under heterogeneous distribution.

Output: Internet data generation resolution matrix H and data reduction set X (A, B) under heterogeneous distribution [9].

(1) If n represents the number of attributes in the internet data decision-making system, then:

$$X(A, B) = \phi \tag{2.2}$$

Among them, X represents the internet data reduction value under heterogeneous distribution, ϕ represents the internet data reduction set under heterogeneous distribution [10].

(2) Assuming an nXn internet data attribute set matrix N;

(3) According to the discernibility matrix in ontology, the data discernibility matrix is generated, and the internet data attribute set matrix N recorded in (2) includes:

$$for(j = i + 2; i < n; j + +) \tag{2.3}$$

$$ifB(x_1), N_{ij} \leftarrow \phi \tag{2.4}$$

Among them, i represents the number of data condition attributes, j represents the number of data decision attributes, and N represents the internet data attribute set matrix.

(4) Add each internet data attribute subset in (2) to XLop (A, B);
(5) Output internet data attribute set matrix N, reduction set XLOP (A, B) [11].

In summary, using the feature selection decision system established above for internet data under heterogeneous distribution, the process of feature extraction and data feature selection for internet data under heterogeneous distribution is completed. In order to improve the quality of data mining in internet data mining under heterogeneous distribution, it is necessary to filter internet data. The optimization method of internet data mining under heterogeneous distribution based on ontology uses information entropy to filter internet data, and the filtering theory value of information entropy is used to input the filtering condition value IT of internet data:

$$IT = (U, A_t, V_x, I_x) \tag{2.5}$$

The output is:

$$IT = (U, A_t, V'_x, I'_x) \tag{2.6}$$

Among them, U represents the set of data attribute values corresponding to each attribute on the internet, V represents the expected information value of internet data mining samples under heterogeneous distribution, V 'represents the expected information value of internet filtered data mining samples, I represents the information function value during the internet data filtering process, I' represents the theoretical value of information entropy during the internet data filtering process, t represents the mean value of internet data attributes, and x represents the internet data attribute value. Sort the attribute values of in ternet data under heterogeneous distribution by making each internet data attribute value $x \epsilon A_t$.

For non internet data attribute values, assuming its attribute values are in an ordered relationship, it can be transformed into data numerical ordering. This step has been optimized in the internet data value sorting process under heterogeneous distribution, so that non in ternet data attributes can also be sorted [12]. After sorting, perform the following steps based on the filtering information function of information entropy for each internet data attribute value $x \epsilon A_t$.

$$fori = 1toK - 1 \tag{2.7}$$

Among them, K represents the maximum specified filtering value for filtering internet data under heterogeneous distribution using information entropy. Using information entropy to filter inter net data under heterogeneous distribution can be defined as:

$$H(C/X; V_1, V_2, ..., V_t) = \sum_{j=1}^{i+1} p(U_j) \sum_{d \epsilon V_D} p(d/U_j) log(d/U_j) \tag{2.8}$$

Among them, H represents the defined value of information entropy for filtering internet data under heterogeneous distribution, and p represents the probability distribution value of internet data attributes. When the amount of internet data under heterogeneous distribution continues to increase and the theoretical value of information entropy data filtering changes little, the optimal filtering value is output to complete the filtering of internet data under heterogeneous distribution. In summary, internet data preprocessing under heterogeneous distribution mainly consists of data feature selection and data filtering. Data preprocessing improves the quality of data mining and reduces data mining time [13,14].

2.2. Internet Data Mining under Heterogeneous Distribution . After completing the data preprocessing of internet data mining under heterogeneous distribution based on cost body theory, decision tree algorithm is used to mine internet data under heterogeneous distribution. The specific methods are as follows: The decision tree algorithm constructs an in ternet data mining decision tree under heterogeneous distribution in a top-down manner, which is divided into internet data decision tree generation and internet data decision tree pruning. The author does not study internet data decision tree pruning. The internet data decision tree generation algorithm utilizes information gain to select the best mining attributes in internet data under heterogeneous distribution. The specific calculation method for information gain is as follows:

Assuming there are m pieces of information, the probability distribution of the data attributes mined is:

$$p = (p_1, p_2, ..., p_m) \tag{2.9}$$

The expected information value of internet data mining sample S under this heterogeneous distribution:

$$V(S) = V(p) = \sum_{1}^{m} p_i log_2 p_i \tag{2.10}$$

Among them, S represents the total number of samples in internet data mining under heterogeneous distribution, and m represents the number of information in the information gain [15]. Given the heterogeneous distribution of internet data mining samples $s_i \epsilon S$, the total number of in ternet data mining samples is s_i, based on the attribute values of internet data mining categories s under the heterogeneous distribution; Divide into z subsets of data category attributes, and the number of internet data mining samples under heterogeneous distribution contained in each subset of data mining categories is s_{ij}. Therefore, the probability distribution of internet data mining attributes is shown in Equation 2.11:

$$p = (S_{i1}/S_i, S_{i2}/S_i, ..., S_{iz}/S_i) \tag{2.11}$$

According to Equation 2.10, the expected information value of internet data mining sample s_i is $I(s_i) = I(p)$. The entropy of the internet data mining sample set S under heterogeneous distribution is:

$$E(S) = \sum_{1}^{z} \frac{S_i 1 + S_i 2 + ... + S_{im} I(s_i)}{S} \tag{2.12}$$

The information gain value of internet data mining sample S under heterogeneous distribution is:

$$Y(S) = I(S) - E(S) \tag{2.13}$$

Among them, Y represents the information gain value of internet data mining under heterogeneous distribution, and E represents the entropy of internet data mining sample set under heterogeneous distribution. Perform iterative calculations on the above process until one of the following conditions is met: all samples of a given internet data node belong to the same classification; There are no additional internet data attributes that can be further divided into data attribute samples; The internet data branch attribute sample under heterogeneous distribution is empty. So far, we have completed the internet data mining under heterogeneous distribution [16,17].

3. Results and Analysis. In order to demonstrate the effectiveness of internet data mining optimization methods based on ontology and heterogeneous distribution, a simulation experiment is required. Build an internet data mining experimental simulation platform under heterogeneous distribution in the environment of Visual C. The experimental data was taken from the SPSS Elementinell. 1 data mining system. In this experiment, ontology was used to high-quality mine internet data under heterogeneous distribution in the SPSS Elementinell. 1 data system. Table 3.1 and Figure 3.1 describe the relationship between the amount of feature selection data (10000) and its selection efficiency (%) in the optimization method of internet data mining under heterogeneous distribution based on text theory.

It is evident from the various data in Table 3.1 that the internet data mining optimization method based on text theory under heterogeneous distribution is safe and reliable. Although the efficiency of data feature selection in the table fluctuates continuously with the increase of feature selection data volume, the selection efficiency is basically above 90%, which further demonstrates the overall effectiveness of internet data mining optimization methods based on text theory under heterogeneous distribution. Table 3.2 and Figure 3.2 describe the relationship between the amount of filtered data (10000 pieces) and the filtering time (s) in the optimization method of internet data mining under heterogeneous distribution based on text theory.

In Table 3.2, the relationship between the amount of filtered data and the time it takes in the optimization method for internet data mining under heterogeneous distribution based on text theory is described. The time

Table 3.1: Relationship between Data Feature Selection and Selection Efficiency in Internet Data Mining

Number of feature selection data (10000)	Data selection efficiency (%)
1000	92.7
2000	93.4
3000	95.3
4000	92.5
5000	94.6

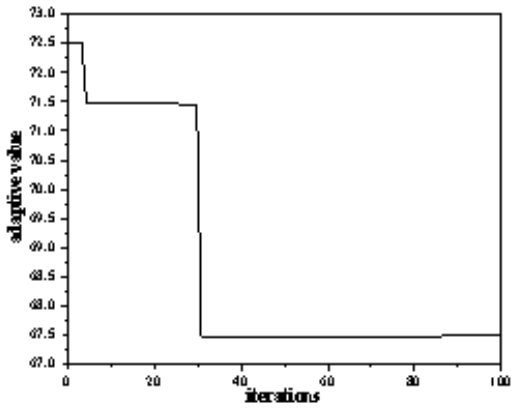

Fig. 3.1: Relationship between Data Feature Selection and Selection Efficiency in Internet Data Mining

Table 3.2: Relationship between Data Filtering and Time Used in Internet Data Mining

Filtered data volume (10000)	Time taken for filtering (s)
100	3.12
200	3.64
300	4.11
400	4.57
500	4.9

it takes to filter data fluctuates relatively little with the increase of filtered data, indicating that the data mining optimization method proposed by the author consumes less time, further proving the feasibility of the optimization method for internet data mining under heterogeneous distribution based on text theory. Figure 4 shows a comparison of the mining efficiency (%) between the fast mining method for hidden data and the author's method. The efficiency of the fast mining method for hidden data in Figure 3.3 fluctuates greatly with the increase of data volume. The mining efficiency of the data mining optimization method proposed by the author is in a stable fluctuation state, and the mining efficiency is relatively high, which is significantly better than the fast hidden data mining method. This is mainly because when using the author's proposed method for data mining on internet data under heterogeneous distribution, the maximum interval algorithm is used for feature extraction on internet data Based on the feature selection decision system of internet data, feature selection is carried out on internet data, as well as the preprocessing work of internet data mining under heterogeneous distribution using information entropy to filter internet data. This has laid a solid foundation for internet data mining under heterogeneous distribution, which is conducive to efficient mining of internet

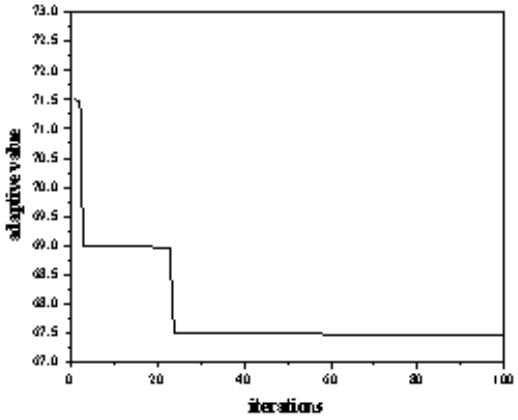

Fig. 3.2: Relationship between data filtering and time used in internet data mining

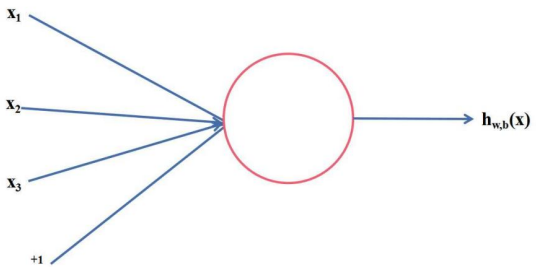

Fig. 3.3: Comparison of Mining Efficiency under Different Methods

data under heterogeneous distribution. Figure 5 shows the comparison of the error rate (%) between the hot topic data mining method and the author's proposed method [18].

The error rate of the internet data mining optimization method based on text theory and heterogeneous distribution proposed by the author in Figure 3.4 is significantly lower than that of the hot topic data mining method. The error rate of the data mining method proposed by the author fluctuates relatively steadily with the continuous increase of the rated number of data mining, and remains below 5%. The main reason is that the generation of internet data decision trees plays an indispensable auxiliary role in the process of data mining, improving the accuracy of internet data mining under heterogeneous distribution, and effectively increasing the feasibility and optimization of the method proposed by the author. Simulation experiments have shown that the optimization method for internet data mining under heterogeneous distribution based on text theory proposed by the author can accurately mine internet data under heterogeneous distribution, ensuring the overall effectiveness of internet data mining, improving the speed of data mining, and providing a reliable basis for research and development in this field [19,20].

4. Conclusion. When using current methods for data mining on the internet under heterogeneous distribution, it is not possible to achieve high-precision and efficient data mining on the internet under heterogeneous distribution, resulting in high mining error rate, slow speed, and insecurity. The author proposes an optimization method for internet data mining under heterogeneous distribution based on text theory. Through simulation experiments, it has been proven that the proposed method can accurately mine in ternet data under heterogeneous distribution, and has good application value and is practical and feasible.

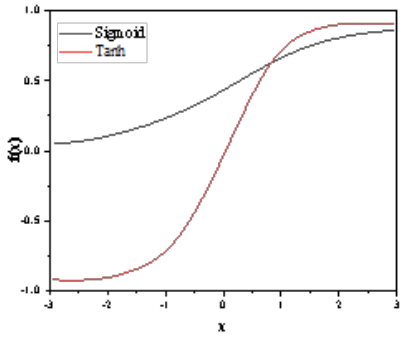

Fig. 3.4: Data Mining Error Rate

5. Acknowledgements. The work was supported by the 2024 Key Research Project Plan for Higher Education institutions in Henan Province; Project number(24B520047)

REFERENCES

[1] Ruifeng, S. (2021). Research on data mining system based on artificial intelligence and improved genetic algorithm. Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology5(4), 40.
[2] Wang, J. (2021). Research on data mining and prediction of large-scale competitions based on online data migration model. Journal of Intelligent and Fuzzy Systems78(2), 1-12.
[3] Hong, Z. , Feng, Y. , Li, Z. , Li, Z. , Hu, B. , & Zhang, Z. , et al. (2022). Performance balance oriented product structure optimization involving heterogeneous uncertainties in intelligent manufacturing with an industrial network. Information Sciences, 598(123), 126-156.
[4] Li, B. S. Z. (2021). Research on data mining equipment for teaching english writing based on application. Journal of intelligent & fuzzy systems: Applications in Engineering and Technology, 40(2),46.
[5] Miao, D. , Lv, Y. , Yu, K. , Liu, L. , & Jiang, J. (2023). Research on coal mine hidden danger analysis and risk early warning technology based on data mining in china. Process Safety and Environmental Protection, 171(245), 1-17.
[6] Zhang, L. , Chen, H. , & Zheng, M. (2022). Research on risk assessment method of energy system based on data mining. International journal of global energy issues46(1), 44.
[7] Wang, Y. W. , Cao, J. G. , Song, C. N. , Wang, L. L. , Sun, L. , & Xie, D. , et al. (2022). Research on high-precision transverse thickness difference control strategy based on data mining in 6-high tandem cold rolling mills. Steel Research International89(6), 93.
[8] Shen, L. (2021). Data mining artificial intelligence technology for college english test framework and performance analysis system. Journal of intelligent & fuzzy systems: Applications in Engineering and Technology, 40(2),134.
[9] Liu, H. (2021). Research on computer simulation big data intelligent collection and analysis system. Journal of Physics: Conference Series, 1802(3), 032052-.
[10] Santos, M. S. , Abreu, P. H. , Fernandez, A. , Luengo, J. , & Santos, J. (2022). The impact of heterogeneous distance functions on missing data imputation and classification performance. Engineering Applications of Artificial Intelligence: The International Journal of Intelligent Real-Time Automation56(111-), 111.
[11] Li, F. , & Gao, W. (2021). Research on the design of intelligent energy efficiency management system for ships based on computer big data platform. Journal of Physics Conference Series, 1744(2), 022026.
[12] Yunita, A. , Santoso, H. B. , & Hasibuan, Z. A. (2021). Research review on big data usage for learning analytics and educational data mining: a way forward to develop an intelligent automation system. Journal of Physics: Conference Series, 1898(1), 012044 (13pp).
[13] Yang, J. , & Liu, Y. (2021). Application of data mining in the evaluation of enterprise lean management effect. Sci. Program., 2021(122), 4774140:1-4774140:13.
[14] Wu, C. , Xia, Y. , Bi, K. , Desjardins, S. , & Lau, D. (2022). Advances in intelligent long-term vibration-based structural health-monitoring systems for bridges:. Advances in Structural Engineering, 25(7), 1413-1430.
[15] Cui, Y. (2021). Intelligent recommendation system based on mathematical modeling in personalized data mining. Mathematical Problems in Engineering, 2021(3), 1-11.
[16] Chen, C. , Feng, T. , Shao, M. , & Yao, B. (2021). Understanding the determinants of spatial-temporal mobility patterns based on multi-source heterogeneous data. Transportation Research Procedia, 52(34), 477-484.
[17] Sun, H. (2021). Intelligent data mining based on market circulation of production factors. Wireless Communications and Mobile Computing, 2021(4), 1-11.

[18] Lin, Y. (2021). Research on the intelligent early warning system for metal mine mining safety. IOP Conference Series: Earth and Environmental Science, 714(2), 022026 (7pp).
[19] Zhang, L. , Yu, W. , Ren, F. , Sun, J. , Liu, X. , & Zhang, N. , et al. (2021). Research on the design of multi-source heterogeneous data application framework for deep sea based on xml. Journal of Physics: Conference Series, 1802(3), 032028 (4pp).
[20] Garcia, J. , Francesc Aguiló, Adrià Asensio, Ester Simó, Marisa Zaragozá, & Masip-Bruin, X. (2021). Data-flow driven optimal tasks distribution for global heterogeneous systems. Future Generation Computer Systems, 125(4647), 792-805.

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Nov 29, 2023
Accepted: Dec 25, 2023