Digital information storage method of power grid enterprises based on random forest

Li Liu

Yunnan Power Grid Co., Ltd.,
Information Center,
Kunming, Yunnan, 650000, China
Email: liliu@mls.sinanet.com

Abstract: The traditional method of information management and storage in power grid enterprises has some problems, such as time-consuming uploading and downloading of files and unclear classification of information. Therefore, this paper proposes a digital information storage method based on random forest for power grid enterprises. After information was collected, Canopy clustering technology was used to clean the data to avoid the interference of repeated data in the classification process of information. Then the random forest algorithm is used to divide the information categories, and then the digital information storage management module is constructed based on the relational database, and the information storage management is completed in a modular way. Experimental results show that the method takes less time to upload and download files, and the highest accuracy of information classification can reach 93.1%, indicating that the method effectively improves the effect of digital information storage.

Keywords: random forest; digital information; data mining; canopy clustering technology.

Reference to this paper should be made as follows: Liu, L. (2022) 'Digital information storage method of power grid enterprises based on random forest', *Int. J. Internet Manufacturing and Services*, Vol. 8, No. 3, pp.243–253.

Biographical notes: Li Liu received her graduate degree from Beijing Union University. Now, she works in Yunnan Power Grid Co., Ltd. Information Center. Her research interests include the digital transformation of electric power, the digitalisation of corporate finance and artificial intelligence.

1 Introduction

With the continuous development of internet technology, the digital information of power grid enterprises is increasing. This kind of information has the characteristics of large density and variety, so it is necessary to effectively manage the information in order to obtain the required information accurately (Nguyen et al., 2021).

In the process of power grid enterprise information management, information collection, information classification, information extraction and information storage will affect the effect of information management. The information to be stored in power grid enterprises includes the operation information of digital power supply system, intelligent dispatching information, market operation information, etc. From the perspective of

demand, it is very important to classify, organise and store the digital information of power grid enterprises according to certain methods (Song et al., 2020). Especially, information management and storage work has great influence on information quality and utilisation efficiency. If information cannot be stored effectively, it is bound to have a certain impact on information application (Yang et al., 2021), such as reducing the efficiency of information extraction and causing information loss. Therefore, it is necessary to conduct in-depth research on digital information storage.

Zhao and Li (2020) and others proposed a data storage method based on density division. In this method, the redundancy of high-density data is reduced by filtering data before data storage, and then the data dimension is reduced by compression processing. Finally, the density partition method is used to store the processed data. Although this method has some flexibility and can divide the density of various types of data, it takes a long time to upload and download files. Lang (2021) proposed a distributed data storage method based on k-distance topology. Firstly, the k-distance topology sub-graph is used to obtain the safe storage location of data, and then appropriate data storage nodes are selected to improve the security and self-protection ability of storage nodes. Although this method has the advantage of high data security, it is easy to confuse the data categories because it does not classify the data effectively. Liu et al. (2019) and others proposed a data security storage method in hybrid cloud mode. Starting from improving the security of data storage, this method encrypts the data according to the ciphertext strategy for the first time before data storage, and then stores the encrypted data in the public cloud; In order to ensure that users can effectively obtain the required data, the data is processed through anonymous key technology and encryption and decryption technology to facilitate users to access the data in the public cloud at any time. Experimental results show that this method can greatly improve the security of data storage, and the encryption and decryption time is short, which will not affect the data application effect, but there is a problem of low data storage efficiency.

However, when facing the problem of digital information storage and management, the above traditional methods do not carry out efficient classification of data information, resulting in unclear classification of information, which consumes more file upload and download time. To solve this problem, this paper takes the digital information of power grid enterprises as the main research object and applies the random forest algorithm to the digital information storage and management, hoping to solve the problems existing in the traditional methods. The specific design ideas are as follows:

1 Collect the digital information of power grid enterprises to be stored by mining message log data, laying a data foundation for subsequent power grid information management.

2 In order to avoid the influence of invalid data on the information storage and management effect, canopy clustering technology was used to cluster the information with similar repetitive characteristics, so as to complete the cleaning of repeated data and avoid the interference of invalid data in the information classification process.

3 Random forest algorithm is used to classify the cleaned data to ensure that the data attributes in each set are relatively independent, which is convenient for subsequent modular management.

4 Build digital information storage management module based on relational database, mainly including relational database support module, relational database support module and mode conversion module, so as to complete the information storage management in a modular way.

5 According to the results of the simulation experiment, the above method can effectively reduce the upload and download time of files, and the maximum time consumption is only about 2.1 s. It can also effectively divide different types of information, and the highest accuracy of information classification is 93.1%.

2 Digital information storage method of power grid enterprises

2.1 Digital information collection of power grid enterprises

Under the background of the increasing number of power users, the amount of power consumption information is increasing, resulting in the aggravation of the task of digital information collection of power grid enterprises. In order to improve the effect of information collection and provide a reliable data basis for enterprise information application, this paper will use data mining technology to collect digital information (Kishani et al., 2019). The data mining method can analyse the equipment information in the power grid through the message log, and then obtain the analysis results to form a data set. When using this method to collect digital information of power grid enterprises, should not only pay attention to the integrity and accuracy of information, but also reduce the end-to-end delay as much as possible to improve the efficiency of information collection (Marsh et al., 2019).

Set the single-hop transmission time of data in the power grid system to t_s, the time for the node to send a single data packet to t_z, and the length of the data queue to l_d, then the transmission time of the data queue is:

$$T_i = (t_s + t_z) \times N \tag{1}$$

If other factors are not considered, the single hop time of the data packet can be expressed by formula (2):

$$D_i = t_s + t_z + T_i (1 - l_d) \tag{2}$$

According to the above formula, to reduce the data mining time, only need to reduce the length of the data queue (Li et al., 2020).

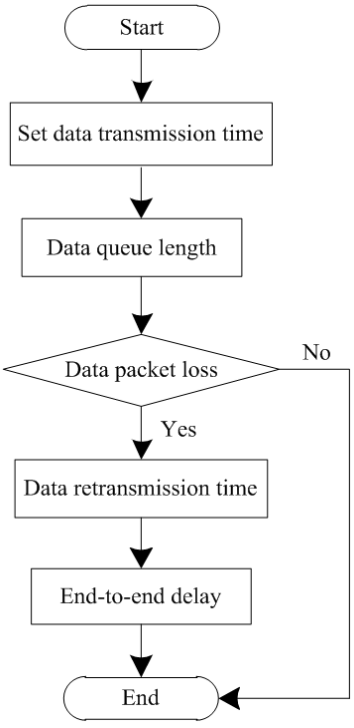
Set the data collection time of a data queue to a fixed value t_k, that is, the value is not affected by the length of the data queue, but when a packet loss event occurs during the data collection process, data retransmission needs to be considered in the overall data mining time. Assuming that the end-to-end delay of the data collection node is t_d, the optimal value of the digital information collection time is:

$$T_j = \sum_{i=1}^{N} \left| D_i + t_k \right|^2 - t_d \tag{3}$$

Among them, N represents the number of data collection nodes.

Figure 1 is the flow chart of digital information collection of power grid enterprises.

Figure 1 Flow chart of digital information collection of power grid enterprises



In Figure 1, on the basis of setting the data transmission time, the length of the data sampling queue is designed, and the end-to-end delay is controlled while data packet loss is controlled, so as to efficiently collect the digital information of power grid enterprises, provide a foundation for power grid information management, and promote the overall improvement of power service level of power enterprises.

2.2 Power grid enterprise digital information data cleaning

After collecting the digital information of power grid enterprises to be stored, further clean the repeated and invalid data to avoid the impact of other data on information storage management (Shanthi and Umamakeswari, 2019).

In order to improve the correctness and effectiveness of information storage, this paper uses canopy clustering technology to clean the digital information, that is, cluster the information with approximate repetition characteristics in the collected data set. First, set a dist function, which is used to determine the degree of repetition between information, (Singh et al., 2019; Tebbi et al., 2019) and its expression is:

$$Dist(x) = \lambda_i \sum_{i=1}^{N} u_i \times |K|^2 \tag{4}$$

Among them, λ_i represents the data source; u_i represents the data dimension; K represents the edit distance, this parameter has a decisive effect on the *Dist* function, and its value directly affects the accuracy of the information clustering degree. The expression is:

$$K = sim(k_i, k_j)V_k \tag{5}$$

Among them, k_i represents valid data; k_j represents repeated data; V_k represents correction weights.

In addition to calculating the value through formula (5), verification is also required to determine the similarity between digital information. In the judgment process, based on the principle of transitivity, the information is clustered and merged to form different cluster sets. If there is a containment relationship between different sets, the included cluster sets can be deleted directly. To eliminate duplicate information and achieve the purpose of digital information cleaning, thereby reducing information redundancy and obtaining high-quality data sets.

2.3 Classification of digital information of power grid enterprises

After data cleaning, the redundant data and invalid data in the data set have been processed. However, in order to achieve efficient storage of the digital information of power grid enterprises, and to enable users to accurately obtain the information they need when extracting information, it is necessary to Perform classification processing to form different data sets to help users make information judgments (Zhou et al., 2021; Mei et al., 2020).

In this paper, random forest algorithm is used to classify the data after cleaning. Random forest algorithm is a statistical classification method, which is based on probability and statistics theory and knowledge classification principle, and has the advantages of accurate classification and not easy to produce misclassification.

According to the characteristics of the digital information of power grid enterprises, according to the characteristics of learning and induction of random forest, a classification structure with tree characteristics, that is, decision tree, is obtained from the digital information resources. Decision tree consists of three core parts, namely decision node, branch and leaf node. According to the different characteristics of digital information of different types of power grid enterprises, the decision points represent the classification characteristics of the digital information to be classified. Branches represent the characteristic dynamic coefficients corresponding to decision nodes. Leaf node represents the digitised information category after classification.

Suppose there is a training sample set $G = \{g_1, g_2, \ldots, g_m\}$, where m represents the number of sample types. If there are M attributes in the set G, the attribute set is expressed as:

$$G_M = \{g_{1m}, g_{2m}, \ldots, g_{mM}\} \tag{6}$$

Classify the sample data in set G. In general, the random forest algorithm can ensure that each attribute in the set is relatively independent. Therefore, the following relationship exists:

$$g(x) = \frac{1}{1+e^x} \tag{7}$$

Among them, e^x represents the data weight.

The data type in the set is judged by the data weight e^x, and the data with the same weight is divided into a set. The specific weight calculation formula is:

$$e^x = X\left(v_h, r_i\right) + \gamma^2 \tag{8}$$

Among them, X represents the number of occurrences of the attribute value; v_h represents the static value of the sample data; r_i represents the dynamic value of the sample data; γ^2 represents the posterior probability.

On this basis, combined with the random forest algorithm, a variety of dynamic characteristics of digital information are added to the data set to prevent the information in the set from being covered by malicious features (Wu et al., 2019). Table 1 shows the classification rules for digital information.

Table 1 Classification rules of digital information

Information associated value	*Specific description*
0	There is a certain similarity in the information, and there is a gap in the weight, which can exist in different data sets at the same time
1	Information is similar, weighted, and exists in the same data set
–1	There is no similarity in the information, the weight difference is large, and it does not exist in the same data set

Combine the weight value with the classification rules, and realise the digital information classification of the power grid enterprise under the condition of considering both at the same time.

2.4 Digital information management and storage of grid enterprises

After digital information collection, information cleaning, and information classification, the preliminary processing of the information is completed. Based on this, a digital information storage management model based on a relational database will be designed. Figure 2 is a diagram of the model composition structure.

The digital information storage and management model based on relational database as shown in Figure 2 mainly includes relational database support module, relational database support module and mode conversion module:

1 Relational database support module

The function of this module is mainly to interpret the instructions issued by the user, then support the data in the relational database, decompose the data into modules that are easy to control and manipulate, and finally process the data and then transmit it to the relational database (Xu et al., 2021a, 2021b).

2 Input/output control module

The input link of the module refers to file upload and the output link refers to file download (Zhou and Xie, 2020). In order to reduce the time consumed by these two links, the two links need to be controlled through the input and output control module to improve the information storage efficiency (Hou et al., 2020). During the operation of the module, it is necessary to understand the basic attributes of information and the characteristics of information types, and sort out the problems to be solved. Sort out the required information, store similar data in the same folder,

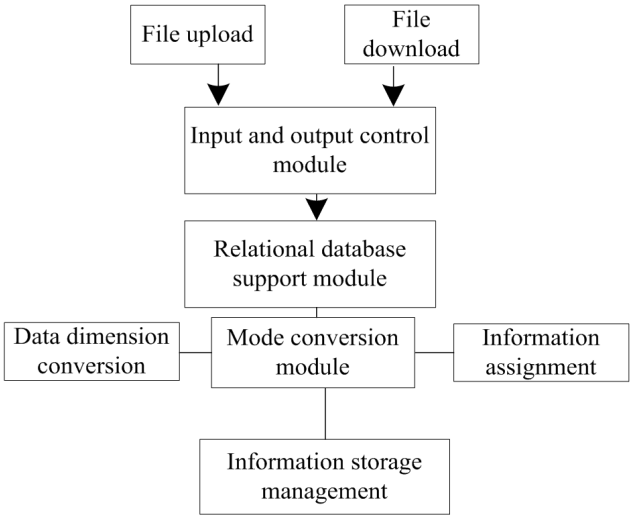
and denoise the data containing noise, so as to achieve the purpose of data dimensionality reduction and improve the accuracy of input and output results.

3 Mode conversion module

The function of this module is to convert the information mode in the database, that is, multi-dimensional and complex information into one-dimensional and simple information, so as to carry out information processing and reduce the difficulty of information processing; At the same time, the module can also describe the relationship of information and assign values to the information in the database, which is more convenient for users to extract the information they need.

Based on the above analysis, the random forest algorithm is used to classify the information, and combined with the relational database management model, the digital information storage management of power grid enterprises is completed.

Figure 2 Digital information storage management model based on relational database



3 Experimental test analysis

In order to verify the comprehensive effectiveness of the random forest-based digital information storage method for power grid enterprises, the following experiments are designed.

3.1 Experimental data set

The experiment was carried out on MATLAB R2020b simulation platform with Intel Core I7-7700CPU and 8GB memory. Zhao and Li's (2020) and Lang's (2021) method were used as comparison methods, and the application effect verification was completed jointly with method of this paper from the perspectives of file upload and download time and information classification effect.

The data used in this experiment were all from Oracle database. The Oracle database contains data indicators such as various types of transaction electricity, transaction electricity price, online electricity, and regional electricity consumption, and can centrally store all kinds of data in business domain, management domain, and network domain. Some data were extracted from the database as experimental sample data. In order to ensure the accuracy of experimental results, data were normalised before the experiment.

3.2 Experimental indicators

1 File upload and download time-consuming: File upload is mainly for the information storage stage, and file download is mainly for the user application stage. The shorter the time-consuming of the two stages, the higher the information processing efficiency.

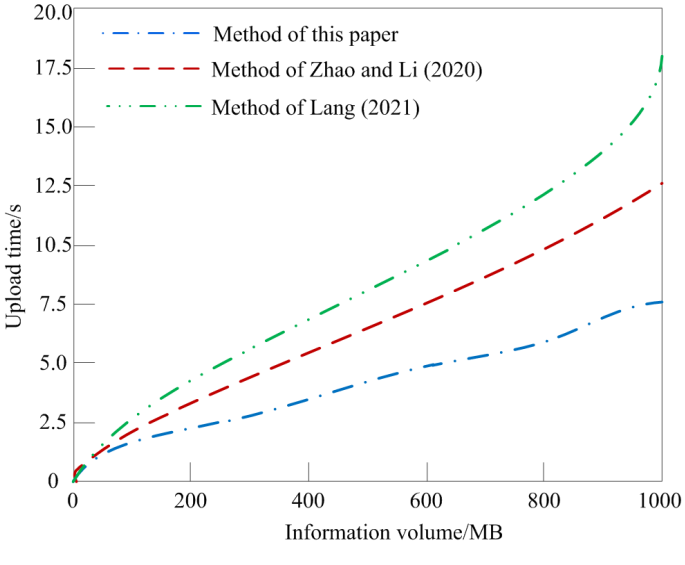
2 Information classification effect: This indicator is mainly reflected by the accuracy of the classification of multiple information types. The higher the accuracy of information classification, the more information users obtain that meets actual needs.

3.3 Experimental results

3.3.1 File uploads and download time-consuming

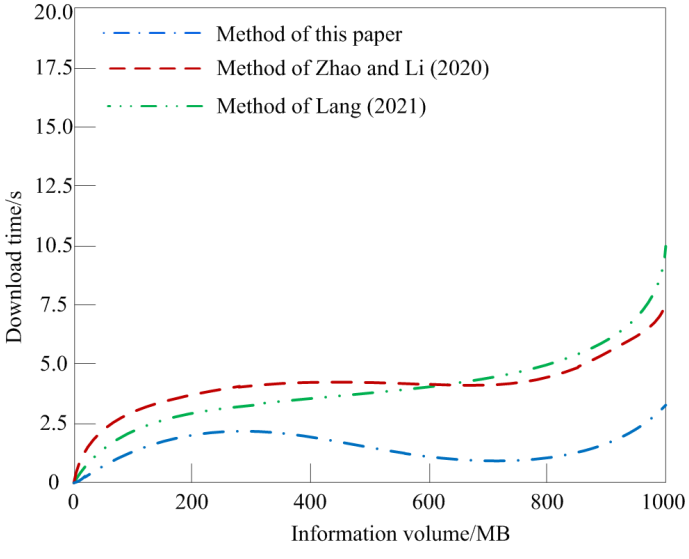
The uploading and downloading of digital information files are time-consuming tests, and the comparison results of different methods are shown in Figure 3.

Figure 3 Time-consuming test results of file upload and download, (a) time-consuming file upload (b) time-consuming file download (see online version for colours)



(a)

Figure 3 Time-consuming test results of file upload and download, (a) time-consuming file upload (b) time-consuming file download (continued) (see online version for colours)



(b)

As can be seen from Figure 3(a), in the process of file uploading, with the increase of information data, the time consumption of the three methods presents a trend of continuous rise, but the rising trend of method of this paper is lower than that of the two traditional methods. The maximum time of file uploading for method of this paper is only about 7.5 s, while that for Zhao and Li's (2020) method is 13.1 s. the maximum time of Lang's (2021) is 18.0 s.

According to Figure 3(b), the time consumption of the three methods fluctuates in the process of file download, but the time consumption of method of this paper is still lower than that of the two traditional methods. The maximum download time for method of this paper is 3.0 seconds.

By comparing the two information processing links of file upload and download, it can be seen that method of this paper can effectively reduce time consumption and has the advantage of high information processing efficiency.

3.3.2 Effect of information classification

Ten types of information were randomly selected from the experimental data set, and the accuracy of the information classification results was used as an indicator to compare the information classification effects of the traditional method and the proposed method. The results are shown in Table 2.

According to the data in Table 2, the accuracy of information classification in this method is higher, with the highest value of 93.1% and the lowest value of 88.4%; the accuracy of information classification in Zhao and Li's (2020) and Lang's (2021) method are slightly lower in the method of this article, the highest values are 89.0% and 91.2%, respectively.

Table 2 Accuracy test results of information classification

Number of experiments/time	*Information classification accuracy rate/%*		
	Method of this paper	*Zhao and Li's (2020) method*	*Lang's (2021) method*
1	89.7	87.4	85.2
2	92.3	82.5	84.9
3	90.5	86.9	83.6
4	88.4	81.9	84.2
5	91.5	88.0	87.1
6	92.3	84.7	86.6
7	90.0	82.2	89.9
8	89.9	85.6	91.2
9	91.6	87.1	90.1
10	93.1	89.0	89.7

The above comparison results show that method of this paper can better classify information types, improve information storage effect, enable users to accurately obtain the information they need, and improve the practicality of information management and storage.

In short, the method of this paper has higher accuracy in classifying information categories and higher efficiency in information processing. This is because method of this paper further classifies digital information on the basis of information collection and cleaning, so that the data set after processing does not contain redundant data and invalid data. Then, the data with the same attributes are normalised according to the weight, which provides the prerequisite for information storage and improves the effectiveness of digital information storage.

4 Conclusions

In view of the problems of existing methods, such as time-consuming file uploading and downloading and unclear information classification, this study proposes a random forest-based digital information management and storage method for power grid enterprises combined with random forest algorithm. The simulation results show that the proposed method can effectively reduce the file upload download time, around its highest takes only 2.1 s, improve the efficiency of information extraction, but also can carry on the effective classification of different types of information, information high classification accuracy is 93.1%, compared with the traditional method, the method of information store the result to be more perfect.

Although the method of this paper has achieved certain application effects, there are still some problems to be optimised. It ignores the real-time variability of digital information. As information may change greatly in a short time, the performance of real-time storage of information still needs to be further studied.

References

Hou, R., Liu, H., Hu, Y. and Zhao, Y.H. (2020) 'Research on secure transmission and storage of energy IoT information based on blockchain', *Peer-to-Peer Networking and Applications*, Vol. 13, No. 4, pp.1225–1235.

Kishani, M., Tahoori, M. and Asadi, H. (2019) 'Dependability analysis of data storage systems in presence of soft errors', *IEEE Transactions on Reliability*, Vol. 68, No. 1, pp.201–215.

Lang, D.H. (2021) 'Distributed data storage method based on K-distance topology', *Journal of Shenyang University of Technology*, Vol. 43, No. 1, pp.67–71.

Li, J., Wu, J. and Jiang, G. (2020) 'Blockchain-based public auditing for big data in cloud storage', *Information Processing & Management*, Vol. 57, No. 6, p.102382.

Liu, X.J., Ye, W., Jiang, J.W. and Zhang, L. (2019) 'Secure data storage scheme in hybrid cloud', *Transactions of Beijing Institute of Technology*, Vol. 39, No. 3, pp.295–303.

Marsh, M., Chaput, T. and Smith, D. (2019) 'Unified electronic traceability and data storage system', *Cytotherapy*, Vol. 21, No. 5, p.43.

Mei, Z., Ding, W., Feng, C. and Shen, L. (2020) 'Identifying commuters based on random forest of smartcard data', *IET Intelligent Transport Systems*, Vol. 14, No. 4, pp.207–212.

Nguyen, T.T., Cai, K., Immink, K. and Han, M.K. (2021) 'Capacity-approaching constrained codes with error correction for DNA-based data storage', *IEEE Transactions on Information Theory*, Vol. 67, No. 8, pp.5602–5613.

Shanthi, P. and Umamakeswari, A. (2019) 'Privacy preserving time efficient access control aware keyword search over encrypted data on cloud storage', *Wireless Personal Communications*, Vol. 109, No. 4, pp.2133–2145.

Singh, A., Garg, S., Kaur, K., Batra, S., Kumar, N. and Raymond, C.K-K. (2019) 'Fuzzy-folded bloom filter-as-a-service for big data storage in the cloud', *IEEE Transactions on Industrial Informatics*, Vol. 15, No. 4, pp.2338–2348.

Song, W., Cai, K. and Immink, K. (2020) 'Sequence-subset distance and coding for error control in DNA-based data storage', *IEEE Transactions on Information Theory*, Vol. 66, No. 10, pp.6048–6065.

Tebbi, A., Chan, T.H. and Sung, C.W. (2019) 'Multi-rack distributed data storage networks', *IEEE Transactions on Information Theory*, Vol. 65, No. 10, pp.6072–6088.

Wu, Q., Zhong, R., Zhao, W., Song, K. and Du, L. (2019) 'Land-cover classification using GF-2 images and airborne lidar data based on random forest', *International journal of remote sensing*, Vol. 40, Nos. 5–6, pp.2410–2426.

Xu, G., Han, S., Bai, Y., Feng, X. and Gan, Y. (2021a) 'Data tag replacement algorithm for data integrity verification in cloud storage', *Computers & Security*, Vol. 103, No. 3, p.102205.

Xu, Z., Wang, Y. and Wang, X. (2021b) 'Research on cloud storage optimization of unstructured big data combined with blockchain', *Computer Simulation*, Vol. 38, No. 7, pp.304–307, p.354.

Yang, C.T., Chen, T.Y., Kristiani, E. and Wu, S.F. (2021) 'The implementation of data storage and analytics platform for big data lake of electricity usage with spark', *The Journal of Supercomputing*, Vol. 77, No. 6, pp.5934–5959.

Zhao, H.Q. and Li, C.L. (2020) 'Data storage method and technology based on density partitioning', *Computer Engineering and Design*, Vol. 41, No. 9, pp.2482–2487.

Zhou, Q. and Xie, J. (2020) 'Mobile client data security storage protocol based on multifactor node evaluation', *Journal of Supercomputing*, Vol. 76, No. 2, pp.1144–1158.

Zhou, W., Yang, H., Xie, L., Li, H. and Yue, T. (2021) 'Hyperspectral inversion of soil heavy metals in three-river source region based on random forest model', *Catena*, Vol. 202, No. 12, p.105222.