Scheduling Knowledge Retrieval Based on Heterogeneous Feature Learning for Byproduct Gas System in Steel Industry

Yangyi Liu* Zhen Lv* Jun Zhao* Ying Liu* Wei Wang*

School of Control Science and Engineering, Dalian University of Technology,
Dalian, P. R. China

(E-mail: lvzheng@dlut.edu.cn).

Abstract: In the steel industry, the scheduling decisions of byproduct gas system are made based on a large number of scheduling rules. It is important to establish an effective retrieval method for scheduling knowledge, for the process of scheduling decision is complex and the amount of scheduling rules is large. In this paper, a retrieval method based on heterogeneous data feature learning is proposed, which could search the rules related to the current system state from a large number of scheduling knowledge, and help to make scheduling decisions. Considering that the data structures between the monitoring data and the scheduling knowledge texts are different, a feature learning method based on convolutional neural network is proposed to extract the key features of the scheduling knowledge, and the full-connected neural network is used to extract the corresponding working condition features from the monitoring data. Due to the features of these two kinds of data are heterogeneous, a correlation analysis model for heterogeneous features based on canonical correlation analysis is constructed, and the features are matched by the matching of maximal canonical correlation. The experimental results showed that the proposed method could effectively extract relevant data features and solve the heterogeneous gaps between the real-time data and the knowledge texts, thus effectively retrieve the corresponding scheduling knowledge in different working conditions providing supports for the scheduling work.

Keywords: Heterogeneous feature learning, knowledge retrieval, convolution neural network, byproduct gas system

1. INTRODUCTION

In the byproduct gas system of steel industry, the scheduling decision is made based on a large number of scheduling rules. The schedulers need to make the scheduling decision based on the real-time monitoring data and the scheduling rules (Zemenkova *et al.*, 2016). The decision process and the large number of scheduling rules are complex. If the schedulers could retrieve the relevant scheduling rules quickly and correctly, the cost time of decision making would be greatly shortened and the decision precision would be improved. This paper realizes retrieving the scheduling knowledge according to the real-time monitoring data, and improve the efficiency and precision of scheduling decisions.

The monitoring data and the texts containing scheduling knowledge have different data structures, they are called heterogeneous data. The difficulty of heterogeneous data retrieval is that the data are represented in different ways. So it is impossible to judge the degree of correlation between the two kinds of heterogeneous data directly, which is called "heterogeneity gap" (Zhuang *et al.*, 2008). Traditional retrieval methods used to be based on keyword matching, but this is not suitable for heterogeneous data retrieval. In order to solve the problem of heterogeneous data retrieval, keyword labels can be added to heterogeneous data. For the data structures of labels is consistent, their correlation degree is easy to judge, thus the problem of heterogeneous data retrieval is solved to a certain degree. However, this method is often time-consuming, and the labels are attached the individuals' subjective wills, which results in uncertainties about the results of the search (Rui *et al.*, 1998).

In the 1990s, scholars proposed a method of heterogeneous data retrieval based on content, which represents the characteristics of heterogeneous data through feature vectors. Then the heterogeneous data retrieval is accomplished by matching the similarity of feature vectors (Zhai *et al.*, 2013). The core of this method is the heterogeneous features and heterogeneous data correlation modeling. In recent years, with the study of machine learning, scholars began to use machine learning algorithms to achieve heterogeneous features and heterogeneous data correlation modeling. The machine learning algorithms can be used to mine the deep correlation features of the heterogeneous data, which makes the quality of the model improve a lot compared with that of the artificial model. For example, Karpathy et al. (Karpathy *et al.*, 2014) used convolution neural network to construct language processing model and Jiang et al. (Jiang *et al.*, 2017) established a deep heterogeneous feature extraction model based on deep belief net.

Machine learning algorithms are suitable to solve the problem of multi-factor or non-linear data modeling. Therefore, based on the characteristics of scheduling knowledge retrieval, a scheduling knowledge retrieval method based on convolution neural network, full-connected neural network and canonical correlation analysis is proposed in this paper (Lv *et al.*, 2015). This method can effectively extract the relevant features of scheduling rule texts by convolution neural network model, extract the relevant features of monitoring data by full-connected neural network. Based on canonical correlation analysis, a heterogeneous features correlation analysis model is constructed, which effectively represents the correlation degree between the two kinds of data, accomplishes the matching between monitoring data and scheduling knowledge, and solves the problem of scheduling knowledge retrieval.

2. PROBLEM DESCRIPTION

In iron and steel enterprises, gas mainly includes blast furnace gas (BFG), coke oven gas (COG) and linz — donawitz process gas (LDG). Taking an iron and steel enterprise as an example, the main gas facilities the factory has is two 300000 m^3 of BFG tanks, one 150000 m^3 of COG tank, three 80000 m^3 of LDG tanks, gas pressurization stations, gas mixing stations, COG refining units BFG releasing towers, COG releasing towers, BFG releasing towers and gas pipe networks (Lv *et al.*, 2018).

In order to give full play to the advantages of the gas scheduling in the allocation of various gas resources, ensure the safe and stable operation of the gas system and reduce the releasing rate of various kinds of gas to the greatest extent, the enterprise has formulated a series of scheduling principles. Under the premise of satisfying the consumption, pressure and quality of the main working procedure, it is necessary to give full play to the function of buffering and stabilizing the pressure of the gas tank. The power station, 130 ton of boiler, starting boiler, hot-rolling mixing station and the synthetic gas conversion mixing station are the main control means, the releasing tower is the security means of the system, and the natural gas from the starting fuel station is the backup gas source. The LDG is used first, the COG is the second, and the BFG is the last to ensure the overall balance of the company's gas system. The regulation scope covers the company's gas generation units, pipe networks, users and various pressurized, release and distribution facilities.

The scheduler adjusts all kinds of gas by means of scheduling rules and historical experience (Lv *et al.*, 2016). However, this scheduling method has strong subjectivity. Due to the difference of personal knowledge and the complexity of field environment, different scheduling decisions may be produced in the same state, so it is difficult to ensure a good and stable scheduling effect. In order to train experienced dispatchers, it is necessary to invest a lot of manpower and material resources.

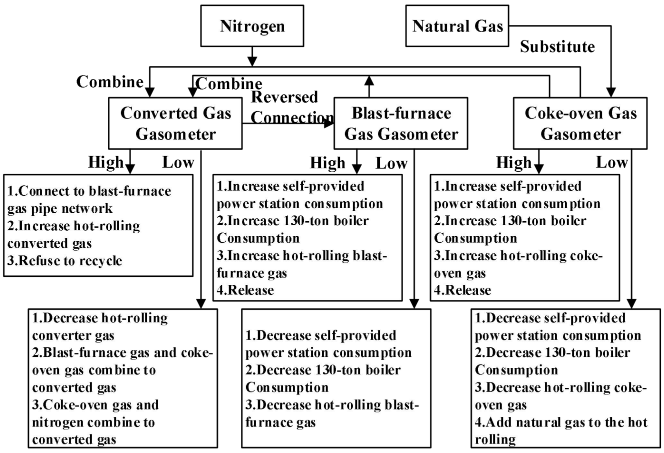
Nitrogen
Natural Gas
Substitute
Combine
Combine
Reversed Connection
Converted Gas Gasometer
Blast-furnace Gas Gasometer
Coke-oven Gas Gasometer
High
Low
High
Low
High
Low
1.Connect to blast-furnace gas pipe network
2.Increase hot-rolling converted gas
3.Refuse to recycle
1.Increase self-provided power station consumption
2.Increase 130-ton boiler Consumption
3.Increase hot-rolling blast-furnace gas
4.Release
1.Increase self-provided power station consumption
2.Increase 130-ton boiler Consumption
3.Increase hot-rolling coke-oven gas
4.Release
1.Decrease hot-rolling converter gas
2.Blast-furnace gas and coke-oven gas combine to converted gas
3.Coke-oven gas and nitrogen combine to converted gas
1.Decrease self-provided power station consumption
2.Decrease 130-ton boiler Consumption
3.Decrease hot-rolling blast-furnace gas
1.Decrease self-provided power station consumption
2.Decrease 130-ton boiler Consumption
3.Decrease hot-rolling coke-oven gas
4.Add natural gas to the hot rolling

Fig. 1. Partial basic scheduling rules.

Scheduling experience means that when facing with a particular system state, the scheduler know which scheduling rule should be followed, and under the guidance of these rules, he can give the adjustment orders to the specific facilities. However, the number of monitoring data is large and the corresponding rules are complex. It is difficult for ordinary schedulers to quickly give the most appropriate scheduling rules for the current system state. If we can learn the correlation between production state data and scheduling rules through artificial intelligence technology, and give the corresponding scheduling rules automatically for real-time monitoring data, it will bring great convenience to the schedulers.

3. KNOWLEDGE RETRIEVAL BASED ON HETEROGENEOUS FEATURE LEARNING

The essential problem of the retrieval of the scheduling knowledge based on the heterogeneous feature learning is obtaining the relevant features of scheduling knowledge and monitoring data by the artificial neural network model, and establishing the degree of correlation between the two kinds of features by means of heterogeneous feature learning (Yuan *et al.*, 2012). Therefore, the purpose of retrieving scheduling knowledge by monitoring data could be accomplished by implementing the mapping of monitoring data to scheduling knowledge texts.

3.1 Text feature extraction based on convolution neural network

The scheduling knowledge involved in scheduling knowledge retrieval usually exists in the form of text. For the text data has the characteristics of large amount and high dimension, in order to grasp the key features contained in the texts and filter independent features, this paper proposes to extract the key features based on convolution neural network as shown in Fig. 2.

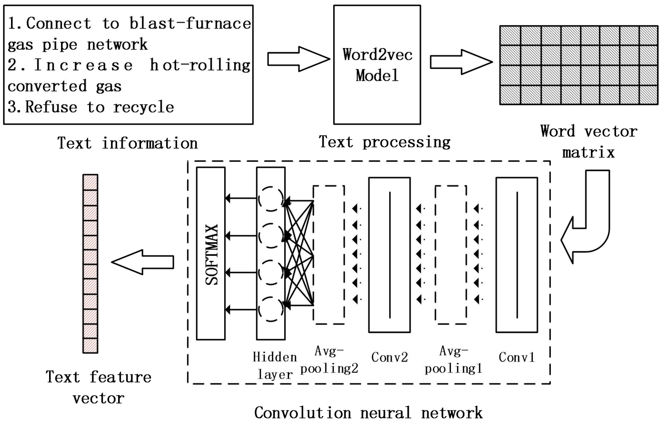

Fig. 2. Text feature extraction model based on convolution neural network.

The input of the feature extraction model is the texts of scheduling knowledge. Texts are converted to word vector matrix after data preprocessing, and is inputted into the convolution neural network. The convolution operation through convolution kernel and input matrix is shown in (1) (2).

$$a_{i,j} = f(\sum_m \sum_n \omega_{m,n} x_{i+m,j+n} + \omega_b) \tag{1}$$

$$f(x) = \max(0, x) \tag{2}$$

Where $x_{i,j}$ represents the element of the line i and column j in the input matrix. $\omega_{m,n}$ represents the weight of the row m and column n in the convolution kernel. ω_b represents the bias term of the convolution kernel. $f(\cdot)$ represents the activation function. $a_{i,j}$ is the element of row i and the column j in the output matrix. In order to reduce the risk of over-fitting in feature extraction and improve the fault tolerance of the model, the average pool layer is added following to the convolution layer to reduce the feature dimension.

$$a^l_{m,n} = \frac{1}{t} \sum a^{l-1}_{i,j} \tag{3}$$

Where $a^{l-1}_{i,j}$ represents the element in the area covered by the pool kernel at the previous layer, t represents the coverage area. $a^l_{m,n}$ represents the element of row m, column n in the output matrix of pool layer. Finally, the extracted features are integrated into the feature vector by the full-connected layer.

3.2 Monitoring data feature extraction based on full-connected neural network

The essence of scheduling is a series of complex logic judgment processes. The basis of make scheduling decision is included by the value of monitoring data. In order to make machine learn the logical relation contained in the monitoring data, the full-connected neural network is trained to learn the judgment method of scheduling based on back propagation algorithm, distinguishing the corresponding working conditions of different monitoring data, outputting feature vectors. The monitoring data feature extraction model is shown in Fig. 3.

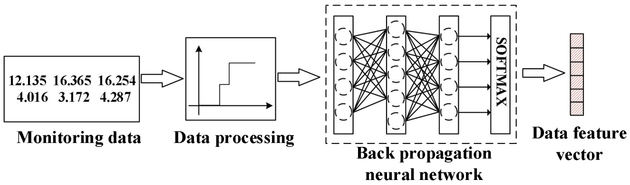

Fig. 3. Monitoring data features extraction model based on full-connected neural network.

In order to distinguish the different states of a monitoring data, a multiple logistic function is used to preprocess the monitoring data. Original logistic function is also called sigmoid function because of its shape of "S". The mathematical expression of sigmoid function is shown in (4).

$$S(x) = \frac{c}{1 + e^{-(x+a)/b}} + d \tag{4}$$

Where a and d are the offset of the transverse coordinate and the offset of the longitudinal coordinate, respectively, b and c decide the shape of sigmoid function. An original sigmoid function can only distinguish two states of an item of monitoring data. Multiple sigmoid function is generated by multiplication of two or more sigmoid functions with different parameters. Multiple sigmoid function can distinguish more than two states. This paper uses the multiple sigmoid function to preprocess the monitoring data to distinguish different states of each item of the monitoring data. After the preprocessing, the data are inputted into full-connected neural network. The outputs of the network get through softmax layer in order to make the differences between each element of the output vector clearer. The mathematical expression of the softmax layer is shown in (5).

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \quad j = 1,\dots,K \tag{5}$$

Where K is the dimension of input vector, z_j is the element of input vector, $\sigma(z)_j$ is the element of output vector. The network model updates the weights of each layer through the back propagation algorithm. Through training neural network, the model accomplishes the purpose of inputting the monitoring data and outputting the feature vector.

3.3 Correlation retrieval method based on canonical correlation analysis

Feature vectors can effectively reflect the internal features of the data, but cannot reflect the relationship between heterogeneous data directly. The purpose of scheduling knowledge retrieval can only be accomplished by representing the degree of correlation between two kinds of data in a certain way (Zhai *et al.*, 2012). Therefore, based on canonical correlation analysis, this paper constructs a analysis model of correlation between scheduling knowledge and monitoring data. Canonical correlation analysis means that two vector groups are compressed into two vectors by linear transformation, and the linear correlation between the two vectors is maximized. The linear correlation between the two vectors represents the canonical correlation between the two vector groups. As shown in (4) (5) (6).

$$\begin{aligned} S_x' &= S_x w_x = \left(w_x^T x_1, w_x^T x_2, ..., w_x^T x_N\right)^T \\ &= (x_1', x_2', ..., x_N')^T \end{aligned} \tag{6}$$

$$\begin{aligned} S_y' &= S_y w_y = \left(w_y^T y_1, w_y^T y_2, ..., w_y^T y_N\right)^T \\ &= (y_1', y_2', ..., y_N')^T \end{aligned} \tag{7}$$

$$\begin{aligned} corr(S_x', S_y') &= corr(S_x w_x, S_y w_y) \\ &= \frac{\operatorname{cov}(S_x w_x, S_y w_y)}{\sqrt{D(S_x w_x)}\sqrt{D(S_y w_y)}} \end{aligned} \tag{8}$$

where S_x and S_y is a pair of matching feature matrices or vector groups of monitoring data and scheduling knowledge texts. The feature matrices is obtained by reshaping the feature vectors. Linear transformation w_x and w_y can be found by optimization calculation in order to maximize $corr(S_x', S_y')$. The maximum of $corr(S_x', S_y')$ represents the canonical correlation between S_x and S_y. By means of canonical correlation analysis, the linear transformation w_{xi} and w_{yi} corresponding to the monitoring data feature matrix S_{xi} and the text feature matrix S_{yi} under working condition $i (i = 1,2,...,n)$ can be obtained. Correlation retrieval method based on canonical correlation analysis is shown in Fig. 4.

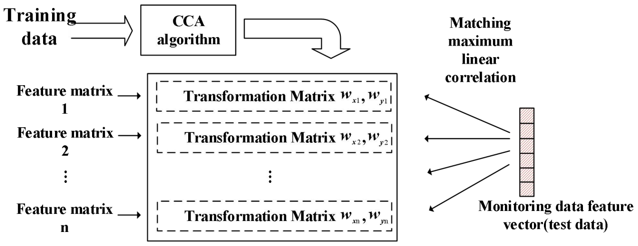
Training data
CCA algorithm
Matching maximum linear correlation
Feature matrix 1
Feature matrix 2
⋮
Feature matrix n
Transformation Matrix w_{x1}, w_{y1}
Transformation Matrix w_{x2}, w_{y2}
⋮
Transformation Matrix w_{xn}, w_{yn}
Monitoring data feature vector(test data)

Fig. 4. Correlation retrieval method based on canonical correlation analysis.

The input monitoring data feature vector is reshaped to feature matrix S_x. Through linear transformation under different working states, S_x is transformed into n kinds of different feature vectors $S_{xi}' (i = 1,2,...,n)$. By calculating the correlations $corr(S_{xi}', S_{yi}')(i = 1,2,...,n)$, taking the maximum correlations as the criterion, the model can find the corresponding working state i_0 and the scheduling knowledge texts. As shown in (9).

$$i_0 = \arg\max(corr(S_{xi}', S_{yi}')) \quad (i = 1,2,...,n) \tag{9}$$

Through the model of correlation analysis between scheduling knowledge and monitoring data, the text data of scheduling knowledge can be retrieved quickly and accurately by monitoring data.

3.4 Scheduling knowledge retrieval based on heterogeneous feature learning

Aiming at the heterogeneity between scheduling knowledge and monitoring data, the heterogeneous feature learning method proposed in this paper can effectively search the scheduling knowledge for the current working condition. The model of knowledge retrieval based on heterogeneous feature learning is shown in Fig. 5, and the processes are listed as following.

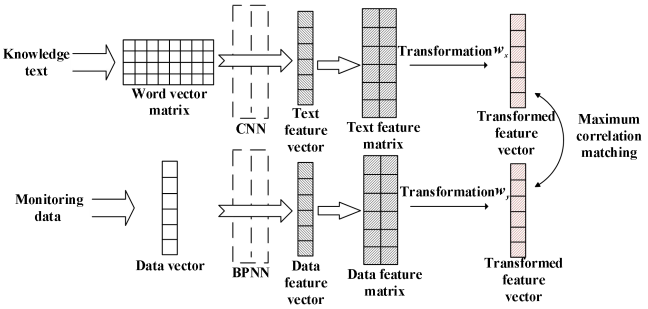

Fig. 5. Scheduling knowledge retrieval based on heterogeneous feature learning.

Step 1. Text processing. Text processing is to extract the texts related to scheduling knowledge, convert each word to a word vector by language processing model word2vec (Ordentlich *et al*., 2016). The word vectors make up the word vector matrices. In order to make the matrices have the same dimensions so as to facilitate the next step, the word vector matrices with lower dimension are processed by complement with zero elements.

Step 2. Scheduling knowledge text feature extraction. The constructed word vector matrices are inputted into the convolution neural network, and their internal features are extracted. The corresponding feature vectors are outputted by the convolution neural network.

Step 3. Monitoring data feature extraction. After the monitoring data is preprocessed, they are inputted into the trained fully-connected neural network. the logic information in the monitoring data extracted by the neural network is to be outputted in the form of feature vectors.

Step 4. Knowledge retrieval. Based on the matched scheduling knowledge texts and field monitoring data, a series of linear transformations are obtained by canonical correlation analysis. Through comparing the correlation of the feature vectors after linear transformed, the scheduling rules whose feature vector has maximal correlation with the monitoring data feature vector is the retrieval target.

4. EXPERIMENTS

In order to verify the effectiveness of the proposed scheduling knowledge retrieval method based on heterogeneous feature learning, the following experiments are carried out. In the experiment, a COG tank, two BFG tanks and three converter gas tank data are used in the experiment. The experiment involves diverse working states, every state has its corresponding scheduling knowledge. According to the field monitoring data of the COG tank, the BFG tank and the LDG tank, the texts of the corresponding scheduling rules is retrieved. In the experiments, monitoring data at each moment only correspond to one kind of scheduling knowledge, so the experimental results will only output one result, and the results would be divided into correct ones and wrong ones. Through many experiments, the proportion of the correct results can be calculated. precision is measured by (10).

$$\delta = \frac{m}{N_a} \times 100\% \tag{10}$$

where N_a represents the time of experiments, m represents the number of correct outputs.

After the feature vectors of the scheduling knowledge and the monitoring data are extracted and the linear transformations are solved by canonical correlation analysis algorithm, the feature vectors are transformed into many different forms. The feature vectors are found by matching the maximum linear correlation. A experiment result of the matching is shown in Fig. 6.

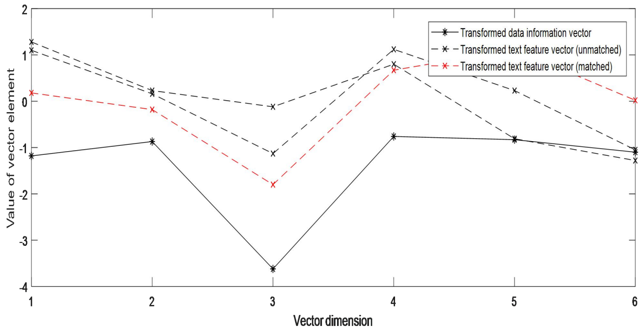

Fig. 6. Feature vectors matching.

For clearer display, some unmatched vectors are discarded from Fig. 6. It can be seen from the graph that the linear correlation between the matched text feature vector and the monitoring data feature vector is the strongest, while the unmatched feature vectors' linear correlations are weak. By calculating the linear correlations, the scheduling knowledge feature vector corresponding to the maximum value are found, and then the target scheduling knowledge is retrieved.

A total of 17980 pairs of samples were used in this experiment. Every sample has a set of monitoring data and a scheduling decision. The samples were used to train the models. In order to demonstrate the effectiveness of the method proposed in this paper, 5000 pairs of samples were randomly extracted in each experiment. After ten experiments, the results of precision δ are shown in Fig. 7.

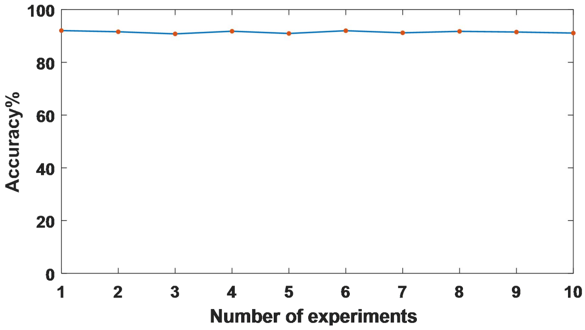

Fig. 7. Ten experiments precision results.

It can be seen from Fig. 7 that, the precision δ of each experiment is above 90%, and the average precision $\overline{\delta} \approx 91.48\%$. Through the analysis, it is found that the main reason that affects the precision of the experiment is that the monitoring data is close to continuous, even after the data processing by multiple logic function, this kind of situation still exists, which leads to the working states of the monitoring data are difficult to judge. Therefore, when the monitoring data is near the critical value of the relevant index, the judgments are often uncertain. In fact, in the real situation, when the monitoring data is near the critical value of the relevant index, the real scheduling rules also have a certain randomness. So no matter the monitoring data are above or below the critical value, either corresponding scheduling rule can satisfy the real situation within a certain range. The results of the experiments may satisfy the real situation. The final effect of this paper is shown in the Fig. 8.

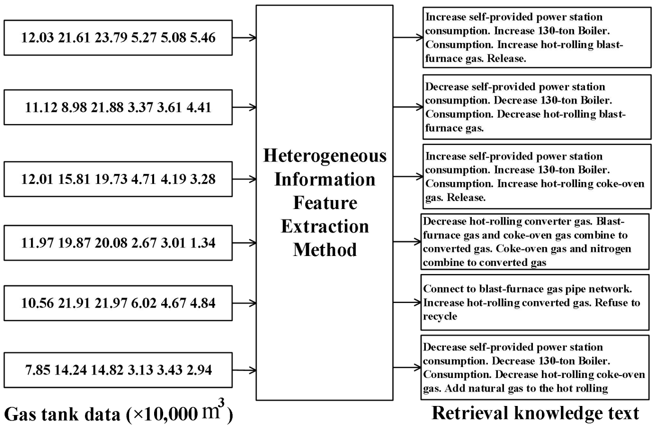
12.03 21.61 23.79 5.27 5.08 5.46
11.12 8.98 21.88 3.37 3.61 4.41
12.01 15.81 19.73 4.71 4.19 3.28
11.97 19.87 20.08 2.67 3.01 1.34
10.56 21.91 21.97 6.02 4.67 4.84
7.85 14.24 14.82 3.13 3.43 2.94
Gas tank data (×10,000 m3)
Heterogeneous Information Feature Extraction Method
Increase self-provided power station consumption. Increase 130-ton Boiler. Consumption. Increase hot-rolling blast-furnace gas. Release.
Decrease self-provided power station consumption. Decrease 130-ton Boiler. Consumption. Decrease hot-rolling blast-furnace gas.
Increase self-provided power station consumption. Increase 130-ton Boiler. Consumption. Increase hot-rolling coke-oven gas. Release.
Decrease hot-rolling converter gas. Blast-furnace gas and coke-oven gas combine to converted gas. Coke-oven gas and nitrogen combine to converted gas
Connect to blast-furnace gas pipe network. Increase hot-rolling converted gas. Refuse to recycle
Decrease self-provided power station consumption. Decrease 130-ton Boiler. Consumption. Decrease hot-rolling coke-oven gas. Add natural gas to the hot rolling
Retrieval knowledge text

Fig. 8. Diagrammatic sketch of scheduling knowledge retrieval.

5. CONCLUSIONS

In this paper, a heterogeneous feature learning method based on convolution neural network and canonical correlation analysis algorithm is proposed, which can be effectively used in the scheduling knowledge retrieval work and provide support to scheduling work. Experimental results have good precision and stability which shows that the constructed convolution neural network and full-connected neural network can effectively extract the relevant features of scheduling texts and field monitoring data and the correlation analysis model effectively reveals the internal relationship between heterogeneous features.

ACKNOWLEDGEMENTS

This work was supported by the National Key R&D Program of China under Grant 2017YFA0700300, the National Natural Sciences Foundation of China under Grant 61703070, Grant 61833003, Grant 61873048, Grant U1908218, the Outstanding Youth Sci-Tech Talent Program of Dalian under Grant 2018RJ01.

REFERENCES

Jiang, B., Yang, J., Lv, Z., Tian, K., Meng, Q., & Yan, Y. (2017). Internet cross-media retrieval based on deep learning. In: *Journal of Visual Communication and Image Representation*, S1047320317300482.

Karpathy, A., Joulin, A., & Fei-Fei, L. (2014). Deep fragment embeddings for bidirectional image sentence mapping. In: *International Conference on Neural Information Processing Systems*. MIT Press.

Lv, Z., Liu, Y., Zhao, J., & Wang, W. (2015). Soft computing for overflow particle size in grinding process based on hybrid case based reasoning. In: *Applied Soft Computing*, **27**, 533-542.

Lv, Z., Zhao, J., Liu, Y., & Wang, W. (2016). Use of a quantile regression based echo state network ensemble for construction of prediction intervals of gas flow in a blast furnace. In: *Control Engineering Practice*, **46**, 94-104.

Lv, Z., Zhao, J., Zhai, Y., & Wang, W. (2018). Non-iterative t – s fuzzy modeling with random hidden-layer structure for bfg pipeline pressure prediction. In: *Control Engineering Practice*, **76**, 96-103.

Ordentlich, E., Yang, L., Feng, A., Cnudde, P., Grbovic, M., & Djuric, N., et al (2016). Network-efficient distributed word2vec training system for large vocabularies.

Rui, Y., Huang, T. S., Ortega, M., & Mehrotra, S. (1998). Relevance feedback: a power tool for interactive content-based image retrieval. In: *IEEE Transactions on Circuits and Systems for Video Technology*, **8(5)**, 644-655.

Yuan, L., Wang, Y., Thompson, P. M., Narayan, V. A., & Ye, J. (2012). Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. *Neuroimage*, **61(3)**, 0-0.

Zemenkova, M., Shalay, V., Zemenkov, Y., Kurushina, E., & Maltseva, T. (2016). Improving the efficiency of administrative decision-making when monitoring reliability and safety of oil and gas equipment. In: *MATEC Web of Conferences*, **73**, 07001.

Zhai, X., Peng, Y., & Xiao, J. (2012). Cross-modality correlation propagation for cross-media retrieval. *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE.

Zhai, X., Peng, Y., & Xiao, J. (2013). Heterogeneous metric learning with joint graph regularization for cross-media retrieval. In: *Twenty-seventh Aaai Conference on Artificial Intelligence*. AAAI Press.

Zhuang, Y. T., Yang, Y., & Wu, F. (2008). Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval. *IEEE Transactions on Multimedia*, **10(2)**, 221-229.