Data Preprocessing and Quality Evaluation for Building the Power Grid Supervision Knowledge Graph

Xinjie Zhang1, *, Jian Wang1, Lingxu Guo2, Xin Wang2, Yao Wang2, and Lei Xu2
1State Grid Tianjin Chengxi Electric Power Supply Branch, 300190, Tianjin, China
2State Grid Tianjin Electric Power Company, 300010 Tianjin, China
* Corresponding author: 910877245@qq.com

Abstract. In the process of constructing the power grid supervision knowledge graph, it is necessary to sort out and integrate structured and unstructured multi-source data for knowledge extraction and reasoning. In order to solve the quality problems of multi-source data redundancy and errors, this paper proposes a multi-source data quality evaluation system to achieve multi-dimensional quality evaluation of power grid supervision data such as maintenance, defect, measurement, alarm and oil chromatography. Data preprocessing methods are firstly adopted for text and numerical data separately to complete data noise reduction, data filtering, data filling, etc. According to the actual calculation example, the practicability and effectiveness of the data preprocessing method and data quality evaluation system are verified. Finally, the data value of the power grid supervision knowledge graph is significantly improved, which is helpful to comprehensively improve the equipment state perception ability.

1. Introduction
With the expansion of the power grid and the complexity of the safe operation environment, the power grid equipment data has several problems such as huge amount of information, overlapped correlation and difficulty in fusion to form knowledge. Also, it lacks real-time assessment of equipment status and trend awareness, so it is urgent to build the power grid supervision knowledge graph, which takes equipment as the core to efficiently store and manage the correlative knowledge. [1-2]
Fig. 1. The flow chart of multi-source data preprocessing and quality evaluation system.

Data sources include structured and unstructured data such as equipment model data, operating data and meteorological data. Considering the original data may exist the problems of messiness, repeatability and ambiguity, it is necessary to carry out data preprocessing and quality evaluation for multi-source data. The standardization of format and content makes data more in line with the needs of knowledge extraction. [3-4] On the other hand, it effectively eliminates redundant data, missing data, uncertain data and inconsistent data to ensure the accuracy and effectiveness of knowledge reasoning.[5-6]

The overall technical route for multi-source data preprocessing and quality evaluation system is shown in Figure 1.

2. Multi-source Data Analysis

The dispatching and control system generates a large amount of data all the time, including alarm information, operating data, and scheduling management data. This paper systematically sorts out the multi-source data needed to construct the power grid supervision knowledge graph. The types, sources and contents of some data are shown in Table 1.

| Data Type          | Data Sources                           | Data Contents                                           |
|--------------------|----------------------------------------|---------------------------------------------------------|
| Equipment Account  | D5000, regulation cloud, PMS           | Equipment name, voltage level, running time, rated capacity, cooling method, etc. |
| Alarm Information  | D5000, regulation cloud                | Alarm threshold, signal, number, over-limit conditions, protection actions, etc. |
According to the data type classification, the power data is divided into structured data and unstructured data.

a. Structured data is highly organized and neatly formatted data. It is a type of data that can be put into tables and spreadsheets. Structured data is also called quantitative data, which is information that can be represented by data or a unified structure. The clear relationship of structured data makes these data very convenient to use, and its main sources are EMS, OMS, regulation cloud and other systems.

b. Unstructured data is data with irregular or incomplete data structure, without a pre-defined data model, and inconvenient to use the two-dimensional logical table of the database to represent the data. Including power grid topology, fault recorder, fault report, image and audio/video information, etc.

3. Multi-source Data Preprocessing

3.1. Text data preprocessing

Natural language processing (NLP) is an important direction in the field of computer science and artificial intelligence. It can be a process of systematic analysis, understanding and information extraction of text data in a smart and efficient way. For non-text data, including power grid topology, fault recorder, fault report, etc., it should be firstly converted into text data. Using NLP technology, through lexical, syntactic, semantic and pragmatic analysis, the conversion of non-text data to text data is realized, which is convenient for subsequent data processing.

The text data preprocessing is mainly to complete the cleaning and standardization. This process will make our data noise-free and directly analyzed. Any text fragment irrelevant to the data context and end-output can be considered to be textual noise. For example, language pause word, URLs or links, special symbols and punctuation marks, etc. This step is to remove all types of noise entities in the text. A common method is to prepare a dictionary of noise entities and then iterate over the text object to remove those tokens that exist in the noise dictionary.

Another type of textual noise is related to multiple expressions of a word. According to the context, if they have similar meanings, the standardization is needed, which convert all the different forms of a word into its canonical form (also called a lemma). The most common practice is stemming and lemmatization. In Figure 2, the standardization of the transformer name is taken as an example, in which “1# transformer” is considered to be a lemma.
Fig. 2. The text standardization of the transformer name.

After the preprocessing of the text data is completed, the standardized data can be used for information association. For example, according to the standardized equipment account, the maintenance and the log are associated. In this way, the empirical information can be transformed into the relationship between the table structure and the system data.

3.2. Numerical data preprocessing

3.2.1. Removing duplicate data
In order to improve the speed and accuracy of data mining, it is necessary to remove duplicate records in the data set. If there are two or more instances representing the same entity, it is a duplicate record. Numerical attributes can be detected by the statistical methods. According to the mean and standard deviation values of different numerical attributes, confidence intervals of different attributes are set to identify and eliminate duplicate records in the data set.

Similarity calculation is often used to obtain the similarity of the record. If the similarity of the two records exceeds a certain threshold, the two records are considered to be matched, otherwise, the two records are considered to point to different entities.

3.2.2. Filling missing data
Filling missing data is another important issue facing the field of data cleaning. Once incomplete and inaccurate data is used for mining, it will affect the correctness of the extraction model and the accuracy of the derived rules. There are currently many methods for missing data filling, which can be divided into two categories:

a. Ignore incomplete data. It can be realized directly by deleting attributes or instances. In the case of small data sets and few incomplete data, this method is often used to achieve data cleaning.

b. Filling incomplete data. In order to fill the missing data, the value closest to it is chosen to replace it to ensure the quantity and quality of the data set. EM (expectation-maximization) algorithm is a common method, in which the maximum likelihood estimate or the maximum posterior estimate of the parameter can be found by creating a probability model. The success of the probability model depends on unobservable hidden variables.

3.2.3. Eliminating Noise data
Noise data refers to data with errors, which normally exist in the data set and may affect the true value of the data set. Common noise filtering methods include regression, mean smoothing, outlier analysis and wavelet method.

a. The regression method is to use a function to fit the data and smooth the data, and then use the smoothed value to replace the original data.

b. The mean smoothing method replaces the original data with the mean value of several adjacent data for variables with sequence characteristics.
c. Outlier analysis firstly uses the clustering method to generate the data sets and put them into clusters. Then the outliers can be detected according to the characteristics of clusters.

4 Multi-source Data Quality Evaluation

4.1 Data quality evaluation system

Five evaluation indexes have been selected to build the data quality evaluation system based on the latest standard “GB/T 36344-2018”, including completeness, accuracy, consistency, timeliness and uniqueness. [7-8]

a. Completeness (Com for short): The completeness includes the completeness of data records and the completeness of data elements, and if the frequency of the time series is incomplete, it is also defined as insufficient completeness, which is shown in Equation 1:

\[ D_{Com} = \frac{\text{Non-empty records}}{\text{Number of all records}} \]  

b. Uniqueness (Unq for short): It is defined by the uniqueness of the attribute value on the data set. The equation is as follows:

\[ D_{Unq} = 1 - \frac{\text{Number of attribute duplicate value}}{\text{Number of all records}} \]  

c. Timeliness (Tim for short): It is calculated based on the characteristics of the time series, and the definition is shown in Equation 3:

\[ D_{Tim} = 1 - \frac{\text{Number of time-series check error}}{\text{Number of all records}} \]  

d. Accuracy (abbreviated as Acc): Considering the occurrence rate of dirty data, data content and data format compliance, the equation is defined as follows:

\[ D_{Acc} = 1 - \frac{\text{Number of outliers}}{\text{Number of all records}} \]  

e. Consistency (referred to as Con): The consistency includes the consistency of the associated data and the consistency of the same data, which is shown in Equation 5:

\[ D_{Con} = \frac{\text{Number of consistency records}}{\text{Number of all records}} \]

4.2 Weight Calculation based on AHP method

Analytic Hierarchy Process (AHP) is used to establish the weight of the five evaluation indexes. The quality fraction is set as the total target layer of AHP, and the above five indexes are set as reference layer, which is shown in Figure 3.

![Fig. 3 Oil chromatographic data quality evaluation model based on AHP method.](attachment:image)

The quality fraction is defined in Equation 6, 7 and 8. \( W \) is the weight vector and \( D \) is judgment matrix of the index fraction.

\[ Q = W \cdot D \]  

\[ W = [w_1, w_2, w_3, w_4, w_5] \]
The Saaty 1-9 scaling method is used to obtain the AHP judgment matrix and the consistency check is required for the solution process. The paper introduces consistency index (C.I.) and random consistency index (R.I.), and examines the judgment matrix \( D \) by judging the consistency ratio (Consistence Ratio, C.R.). The equation 9 is as follows:

\[
C.I. = \frac{\lambda_{max} - n}{n-1} \tag{9}
\]

\[
C.R. = \frac{C.I.}{R.I.}
\]

Thus, \( \lambda_{max} \) is the maximum eigenvalue of the judgment matrix \( D \), \( R.I. \) can be obtained by checking the random consistency index table. When \( C.R. < 0.1 \), \( D \) satisfies the consistency check, \( W \) is the weight vector of the corresponding indexes. Otherwise, the judgment matrix \( D \) needs to be reconstructed until it passes the consistency check.

The three-year oil chromatogram data of 220kV and above transformers in a certain province is taken to obtain the AHP judgment matrix \( D \) which is shown in Equation 10:

\[
D = \begin{bmatrix}
1 & 2 & 4 & 5 & 6 \\
\frac{1}{2} & 1 & 3 & 4 & 5 \\
\frac{1}{4} & \frac{1}{3} & 1 & 3 & 2 \\
\frac{1}{5} & \frac{1}{4} & \frac{1}{3} & 1 & 2 \\
\frac{1}{6} & \frac{1}{5} & \frac{1}{2} & \frac{1}{2} & 1
\end{bmatrix}
\tag{10}
\]

According to the above equations, \( C.I. = 0.0200 \), \( C.R. = 0.0179 < 0.1 \), the judgment matrix \( D \) satisfies the consistency check. Then the weight matrix is obtained, which is shown in Equation 11.

\[
W = [0.4452, 0.2922, 0.1333, 0.0765, 0.0528]
\tag{11}
\]

The availability of data quality determines the efficiency of the power grid supervision knowledge graph. Only the data with a rating of good or above can support the subsequent data analysis work, while that with a rating of average and below needs to be checked and reprocessed immediately. The classification of data quality rating is shown in Table 2.

| Index    | Weight |
|----------|--------|
| 0≤Q≤0.5  | bad    |
| 0.5<Q≤0.75 | average |
| 0.75<Q≤0.9 | good   |
| 0.9<Q≤1   | excellent |

5. Case Analysis and Result

Take the annual oil chromatographic online data of a 220kV transformer of some substation in 2020 as an example, data preprocessing and quality evaluation system is applied to score the data set, the preprocessed Oil chromatogram data set is shown in Figure 4.
The sensor sampling period of this transformer is 2 hours, and it is resampled at an interval of 4 hours. After resampling, there are 1643 records in total. Perform analysis of different dimensions on the resampled time series.

1) Completeness: the missing rate is 11.6%, the completeness score $Com = 0.884$;
2) Timeliness: if there are 11 values of time-series monotonicity that are not satisfied, the timeliness score $Tim = 0.993$;
3) Uniqueness: If the repeat value is 0, the uniqueness score $Unq = 1$;
4) Accuracy: The proportion of outliers is 0.025, and the accuracy score $Acc = 0.975$;
5) Consistency: Consistency score $Con = 1$.

In summary, with the combination of the data quality weights and index score, the total score of the oil chromatogram data set is 0.940, which means the quality rating is excellent and the value for subsequent knowledge extraction and knowledge reasoning is greater.

Fig. 4. Oil chromatogram data set after data preprocessing.

6. Conclusions
This paper mainly conducts in-depth research on the quality evaluation and preprocessing methods for multi-source data. In order to build the power grid supervision knowledge graph, data preprocessing methods are applied to deal with multi-source data such as defect, maintenance, measurement and alarm data. Text data preprocessing is mainly to complete the cleaning and standardization, and numerical data preprocessing is to solve problems of duplicate data, missing data and noise data. It can improve the accuracy and availability of equipment supervision data. Then the quality evaluation system is established to evaluate the preprocessed multi-source data. Utilizing the optimized multi-source data, the dynamically updated power grid supervision knowledge graph can be constructed with high accuracy and effectiveness.
ACKNOWLEDGEMENT
As a stage achievement, this work was only supported by Science and Technology Project of State Grid Tianjin Electric Power Company (No. KJ21-1-6).

References
[1] T. Pu, Y. Tan, G. Peng, H. Xu, Z. Zhang (2021) Construction and Application of Knowledge Graph in the Electric Power Field. J. Power System Technology, 45(6):2080-2091.
[2] H. E, W. Zhang, S. Xiao, et al. (2019) Survey of Entity Relationship Extraction Based on Deep Learning. J. Journal of Software, 30(6):1793-1818.
[3] B. Li, D. Chen, X. Sun, et al. (2021) Multi-Feature Matching Search Algorithm for Bug Knowledge. J. Acta Electronica Sinica, 49(4):661-664.
[4] X. Zhang, J. Wang (2019) A Method for Generating and Checking D5000 Alarm Information Based on Semantic Analysis. J. Electric Power, 52(5):134-141.
[5] Q. Kong, C. Ye, Y. Sun (2018) Research on Data Preprocessing Methods for Big Data. J. Computer Technology and Development, 28(5):1-4.
[6] D. Deng, Y. Xu, J. Chen, R. Yang (2019) Acquisition and preprocessing of smart electric appliance network power data. J. Power Systems and Big Data, 22(3):81-86.
[7] B. Wang, B. Fang, Y. Wang, et al. (2016) Power System Transient Stability Assessment Based on Big Data and the Core Vector Machine. J. IEEE Transactions on Smart Grid, 7(5):2561-2570.
[8] X. Zhang, J. Wang, W. Wang, et al. (2018) The Application and Assessment of Smart Technologies for Batch Control of Load Dispatching. In: 2018 China International Conference on Electricity Distribution. Tianjin. pp:2705-2709.