Research on comprehensive evaluation of data link based on G1 method and entropy weight method

Ming Jin, Junyan Zhang, Shuo Cui, Mei Kang, Yi Xiao, Rongxin Xiang, Zhenye Yan
1Big Data Center of State Grid Corporation of China, Xicheng District, Beijing, China
2Big Data Center of State Grid Corporation of China, Xicheng District, Beijing, China
3Big Data Center of State Grid Corporation of China, Xicheng District, Beijing, China
4Big Data Center of State Grid Corporation of China, Xicheng District, Beijing, China
5Big Data Center of State Grid Corporation of China, Xicheng District, Beijing, China
6School of Control and Computer Engineering, North China Electric Power University, Changping District, Beijing, China
7School of Control and Computer Engineering, North China Electric Power University, Changping District, Beijing, China

Abstract. With the rapid development of mobile Internet technology and the advent of the 21st century Internet and information age, data has become the most important strategic core resources. With the rapid development of ubiquitous power Internet of things construction, there are still many problems in the data link monitoring and data management: the low effectiveness of data transmission affects the normal operation of the data link; there are few data evaluation models, which lead to the problem of one sidedness and subjectivity in the understanding of the operation status of the data link. Aiming at the problem that the current data link cannot directly show the link quality, this position proposes a comprehensive evaluation method for the entire data link based on the order relationship analysis (G1) method and the entropy weight method. First, establish a complete data link indicator system, and then separately use the order relationship analysis method to determine the subjective weight, and use the entropy weight method to determine the objective weight, and finally combine the two to determine the comprehensive weight, and obtain a comprehensive score that reflects both the expert level and experience and the actual situation related data. The data and transmission conditions of each link are shown, and the comprehensiveness, rationality and scientificity of the method are also demonstrated.

1. Introduction
At present, the data in the power industry is gradually increasing. The quality of power data is related to all aspects of people's lives. Data runs through all aspects of the construction of the ubiquitous power Internet of Things, and is the key to the construction of the ubiquitous power Internet of Things. At present, China has made great development and breakthrough in the full link data analysis of big
data interaction, but there are still some problems that are difficult to solve, such as accurate data acquisition, full link data storage, and diversified processing of big data types [1]. In the process of data transmission, the data link is the main undertaker, so the monitoring and evaluation of the data link status has become an urgent problem to be solved. At present, the relevant departments of the power industry have realized the monitoring of the horizontal and vertical links of power data by constructing a two-level data center. Nowadays, the research of comprehensive evaluation by constructing a complete index system is applied in various fields, and comprehensive evaluation is obtained through a variety of methods for specific goals. Comprehensive evaluation is aimed at the research object, establishing an evaluation index system, using a certain method or model to analyze the collected data, and make a quantitative overall judgment on the things being evaluated [2]. This kind of evaluation method has the advantages of comprehensiveness and system, and effectively avoids the limitations of single method evaluation. Now combined with the comprehensive evaluation of the entire data link based on the order relationship analysis method (G1) and the entropy weight method, the data link with the problem can be found in time, effectively improving the efficiency of data link maintenance and data resource monitoring, and ensuring data Link efficient and high-quality circulation.

2. Data link introduction

In the current background of big data in electric power, the problem of data transmission has become more and more important. The data link plays a very important role in the data transmission process. (The definition of data link in this article is no longer the data link layer that provides services to the network layer in the traditional OSI reference model) The data link is mainly composed of three parts: Oracle GoldenGate, two-level data center, and data through channel. The overall structure is shown in the figure below.

![Data link overall architecture](image)

2.1. Oracle GoldenGate

Oracle GoldenGate is a log-based structured data replication software. It mainly obtains incremental changes in data by analyzing the logs in the source database, and then applies the changes to the target database in a specific way, thereby synchronizing data from the source to the target, enabling data to be synchronized quickly and effectively. This separate architecture enables each module (capture, routing, transformation, and delivery) to execute independently their own tasks, to achieve high performance, high reliability and flexibility. It has data compression and encryption functions, reducing the pressure on data transmission bandwidth, and greatly improving the security of data transmission [3]. Ogg also has the function of breakpoint transmission. Ogg generates a new file format trail file. The trail file can prevent single point of failure, persist the transaction information, and have
a checkpoint mechanism to record the current read-write position. In case of failure, the data can be transmitted accurately again through the location of checkpoint record \cite{4}. Ogg supports synchronization under heterogeneous operating systems and heterogeneous databases, and supports mainstream operating systems and database platforms \cite{5}. Its specific structure is shown in the figure below.

![OGG overall architecture](image)

Figure 2 OGG overall architecture

OGG can realize short-term real-time replication of massive data between most operating platforms and database platforms, so it can be applied in multiple scenarios such as data synchronization, disaster recovery, and real-time data tracking. In the data link, OGG mainly plays the role of data synchronization. The specific process is: the source OGG program configures the database tables that need to be synchronized, collects data regularly through the Extract process, and writes the collected data to the trail of the local disk. File, and then deliver the trail file data to the target through the Pump process. The Collector process on the target side receives the data from the source side and generates the Trail file. The Replicat process is responsible for parsing the Trail file and analyzing the results.

2.2. Two-level data center

The data center refers to a platform that can effectively realize data division, horizontal decoupling, and effectively deposit public data by using big data technology and various professional tools and software. It optimizes integration and knowledge accumulation of data in different fields by using modeling methods, and relies on data services to complete encapsulated data services \cite{6}. At present, the data center can be roughly divided into three categories: the first is the data center proposed by Alibaba, the second is the business center, and the third is the fragmented center. The two-level middle station in the data link adopts Ali middle station, which adopts both offline monitoring and real-time monitoring, and each performs its duties and complements each other to build a middle station monitoring capability system. The middle station is mainly composed of MaxCompute, DataWorks, DI, DataHub, RDS for Mysql, DRDS, ADS, DTS and other components. Through the data center, you can fully understand the operating status of the data link two levels and the data through link, including the access status of each business system, data storage and update status, data link status, data quality status, data security status, etc. Important indicators.

The data penetration channel is mainly based on the mainstream tools of the power industry, combined with the integration of data middle station access, to achieve a complete two-level data connection and two-level middle station connection link monitoring. Through the data through channel, it is possible to collect and analyze the two-level data in the Taizhong data storage and data link information.
3. Method introduction

3.1. Entropy method

In information theory, entropy is a measure of the degree of disorder of a system, and it can also measure the effective information provided by data \cite{7}. The concept of "information entropy" is proposed by Shannon by borrowing the meaning of entropy in thermodynamics, and it is used to describe the uncertainty of information sources. Entropy method is to use the actual value of the required evaluation index to measure its effective information, so as to determine the corresponding weight of the evaluation index \cite{8}. The greater the entropy, the smaller the system information and the smaller the index weight; on the contrary, the more certain the system is, the greater the amount of information and the greater the index weight \cite{9}. Therefore, information entropy can be used to calculate the weight of each index, which provides a basis for the comprehensive evaluation of multiple indicators.

The specific calculation steps using the entropy method are as follows:

1) Perform data preprocessing on the original data matrix, generally using the Min-Max standardization method. Suppose m evaluation indicators, the original data matrix obtained by n evaluation objects is

\[
X = \begin{pmatrix}
  x_{11} & x_{12} & \ldots & x_{1n} \\
  x_{21} & x_{22} & \ldots & x_{2n} \\
  \vdots & \ddots & \ddots & \vdots \\
  x_{m1} & x_{m2} & \ldots & x_{mn}
\end{pmatrix}
\]  

(1)

Use the Min-Max standardization method to standardize the matrix to obtain the following matrix

\[
R = (r_{ij})_{m \times n}
\]  

(2)

Where \( r_{ij} \) is the standard value of the j-th evaluation object on the i-th evaluation index, \( r_{ij} \in [0,1] \). For the larger index, there is

\[
r_{ij} = \frac{x_{ij} - \min\{x_{ij}\}}{\max\{x_{ij}\} - \min\{x_{ij}\}}
\]  

(3)

Correspondingly, for the index that the smaller is better, there is

\[
r_{ij} = \frac{\max\{x_{ij}\} - x_{ij}}{\max\{x_{ij}\} - \min\{x_{ij}\}}
\]  

(4)

2) Calculate the entropy value of each index. In an evaluation index system with m evaluation indexes and n evaluation objects, the entropy value of the i-th index is defined as

\[
H_i = -k \sum_{j=1}^{n} p_{ij} \ln p_{ij}, \quad i = 1,2, \ldots, m
\]  

(5)

Where \( p_{ij} = r_{ij} / \sum_{j=1}^{n} r_{ij}, \quad k = 1/ \ln n \). When \( p_{ij} = 0 \), let \( p_{ij} \ln p_{ij} = 0 \).

3) Calculate the weight of each indicator based on the entropy value of each indicator. The calculation method of the weight of the i-th indicator is as follows:

\[
\omega_i = \frac{1 - H_i}{m - \sum_{i=1}^{m} H_i}
\]  

(6)

3.2. Order relation analysis method(G1)

The order relationship analysis method composes the index into the order relationship in the index set \( \{x_1,x_2,\ldots,x_m\} \), that is
Among them, \( x^*_m \) represents the single factor index after forming the order relationship, and the importance ratio \( r_k \) between the evaluation index \( x_{k-1} \) and \( x_k \) is given (i.e. \( \omega_{k-1}^*/\omega_k^* \)), and the value of \( r_k \) is referenced as shown in Table 1. After all the indicators are sorted according to the Table and the importance degree ratio \( r_k \) is determined, the weight coefficient \( \omega_k^* \) corresponding to each indicator can be calculated. The specific calculation method is shown in the following two formulas.

\[
\omega_k^* = \left( 1 + \sum_{i=2}^{m} \prod_{1=k}^{m} r_i \right)^{-1}
\]

(8)

\[
\omega_{k-1}^* = r_k \omega_k^*, k = m, m-1, ..., 3, 2
\]

(9)

| \( r_k \) | Description |
|---------|------------|
| 1.0     | The index \( x_{k-1} \) is as important as \( x_k \) |
| 1.2     | The index \( x_{k-1} \) is slightly more important than \( x_k \) |
| 1.4     | The index \( x_{k-1} \) is obviously more important than \( x_k \) |
| 1.6     | The index \( x_{k-1} \) is strongly more important than \( x_k \) |
| 1.8     | The index \( x_{k-1} \) is extremely important than \( x_k \) |

Table 1 Reference table for relationship assignment

4. Experiments and results

4.1. Data Sources

In this experiment, each link is scored through a comprehensive evaluation model based on various data indicators in the data link, and a comprehensive score for each data link is obtained. The experimental data is mainly collected by the State Grid Corporation of the real-time data of the entire link of the power data, and part of the link failure data is obtained through the simulation software, enriching the data types, and gradually analyzing each data to construct an index system to obtain a comprehensive comprehensive evaluation score for each data link , To quickly and effectively pay attention to the problematic link.

4.2. A subsection

The method of scaling the data according to a certain ratio so that the data can be within a specific interval is called data standardization. In this experimental model, because the difference between different indicators is large, and each indicator data has Different dimensions and orders of magnitude have a greater impact on model evaluation. Therefore, data normalization is required to simplify calculations. The specific steps are

1. Convert data index levels, merge data items with the same index, and set specific index parameters.
2. The data is preprocessed by max-min, the original data is linearly transformed, and the data is uniformly mapped to the interval [0,1]. If the data link index is small or medium, the conversion function is

\[
r_{ij} = \frac{\max\{x_{ij}\}_j - x_{ij}}{\max\{x_{ij}\}_j - \min\{x_{ij}\}_j}
\]

That is, subtract the index value from the maximum value in the index value class, and then do a quotation with the difference between the maximum value and the minimum value in this type of index.

Where max is the maximum value in the sample data, and min is the minimum value in the sample data.
After the data is normalized, the calculation accuracy of the evaluation model in this experiment can be effectively improved. This article mainly uses the above methods to process data.

4.3. Data evaluation rules
According to the comprehensive evaluation rules, the indexes in the index system for comprehensive evaluation of the data full link abnormality are first standardized. Use the entropy method to obtain the entropy value of the standardized index system. After obtaining the entropy value of each secondary indicator, the weight of each secondary indicator can be determined according to the entropy value, and then the standardized indicator value is the same as the corresponding weight Multiply and then sum to get an objective comprehensive evaluation score.

The order relation method is used to obtain the subjective weight vector and the subjective comprehensive evaluation score according to the standardized index set. Combining the entropy weight method and the order relationship analysis method, the Lagrangian multiplier method is used to optimize the calculation of the comprehensive weight, and the final comprehensive evaluation score is obtained. The score can not only reflect the subjective expert opinions, but also fully consider the objectivity and reasonableness of the data Sex.

4.4. Experimental results
Based on the real-time link data, the evaluation results are obtained, and the entropy weight method score, the order relation method score, and the final comprehensive score are calculated respectively. This table can clearly reflect the current real-time data link operation, and can accurately and timely find abnormal links.

| Link serial number | Score by order relation method | Score by entropy weight method | Comprehensive score |
|--------------------|--------------------------------|-------------------------------|--------------------|
| 1                  | 73.0972000708843               | 68.3287607620465              | 70.2361364855816   |
| 2                  | 71.4280761191516               | 68.3784698315755              | 69.598312346606    |
| 3                  | 73.8632295023272              | 68.4760195853467              | 70.6309035521389   |
| 4                  | 73.1087580182975             | 72.2580183436797              | 72.5983142135268   |
| 5                  | 83.9535464535465           | 84.5126247144649              | 84.2889934100975   |

5. Conclusion
This project is based on the data itself, by collecting real-time data indicators and constructing a variety of evaluation method models to evaluate the operation of real-time data links, and propose a data link evaluation method based on entropy weight method and G1 method.

This article introduces the importance of data in the current era of big data and the necessity of data links in the power industry. It also discusses various methods for comprehensive evaluation and the feasibility of applying them to data links.

The evaluation method realized in this paper is mainly based on the entropy weight method and the order relationship analysis method. Based on the subjective weighting of the G1 method, the weight value of the evaluation index obtained is subjectively reasonable, but the accompanying subjective arbitrariness is relatively large and cannot reflect the objective New changes in the premise; based on the objective weighting of the entropy method, the in-depth exploration of the information carried by the original data, but it does not take into account the expert experience and the tendency of decision makers, and the weight value obtained may not be consistent with the actual situation. Therefore, after calculating the subjective weight value and the objective weight value separately, and combining the two, there are both subjective evaluation methods that can reflect the level and experience of experts, and objective evaluation methods that reflect relevant data of the actual situation.

Finally, the experimental results show the scoring of each real-time data link using the method in this article, showing the reliability and feasibility of the method.
References

[1] Peng Zhihui. Research on theoretical system of public security information science under the environment of big data [J]. Library Journal, 2018, 38 (2): 11-17

[2] Jia Junping, he Xiaoqun, Jin Yong. Statistics (4th Edition): China Renmin University Press, 2009: 427

[3] WEI Ming-jun, WU Qian, CHEN Hong-li. Implementation of Blood Station Cloud Disaster Recovery System Based on Oracle GoldenGate [J]. Information and Communication, 2017(10):50-52.

[4] SONG Hu, ZHANG Ming-ming, HU Jun-jun, et al. Research on the Separation Technology of Reading and Writing Database on the Disaster Recovery End [J]. Information Technology, 2017(01):26-29, 33.

[5] CAI Bo, SHAN Chuan, QIAN Kun. Application of Real-time Database Software in Data Exchange Platform System of Hydropower Station [J]. Hydropower Automation and Dam Monitoring, 2014, 38(02):5-9.

[6] Deng Lijun. Analysis of data center and big data center [J]. Electronic world, 2019 (22): 85-86

[7] Meng Qingsheng. Information theory [M]. Xi'an: Xi'an Jiaotong University Press, 1989:19-36

[8] YANG H., GUO M., XIA Y., et al. Trajectory tracking for wheeled mobile robots via model predictive control with softening constraints [J]. IET Control Theory & Applications, 2018, 12:206－214.

[9] GUOJ H., LUO Y G., LI K Q. Integrated adaptive dynamic surface car-following control for nonholonomic autonomous electric vehicle [J]. Science China Technology Sciences, 2017, 60(8):1221－1230.