Research on Electric Power Data Governance System and Its Application

Chen Peng
State Grid Ningxia Electric Power Co.LTD
State Grid Corporation of China
Ningxia, China
745409549@qq.com

Cai Bing
State Grid Ningxia Electric Power Co.LTD
State Grid Corporation of China
Ningxia, China
1174032832@qq.com

Cheng Yi
School of Control and Computer Engineering
North China Electric Power University
Beijing, China
cy_ncepu@163.com

Hu Yang*
School of Control and Computer Engineering
North China Electric Power University-
Beijing, China
* Corresponding author:
hooyoung@ncepu.edu.cn

2021 International Conference on Power System Technology (POWERCON) | 978-1-6654-0737-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/POWERCON53785.2021.9697771

***Abstract*—With the advent of the era of the data economy, the massive data assets accumulated in power companies will become a key element of new strategies for power security and the creation of new kinetic energy for the development of electric power companies. This paper proposes electric power data governance system based on the data quality problems in the electric power data, that is, the power data is divided into measurement and archive data, and it is perfected through five steps of business understanding, data understanding, data preparation, model establishment and data correction. The basic framework of data governance, combined with the analysis and application of the governance system of PMS wire archive data, verifies the practicability and effectiveness of the proposed framework.**

Keywords—electric power data, data governance, PMS wire data

I. INTRODUCTION

With the continuous development of modern smart grids, the informatization and intelligent development of the power system has been promoted. There are many historical data, condition monitoring data and electric energy measurement data in the operation process of power system, and the growth rate is relatively fast, which fully shows the characteristics of the data. Therefore, with the increasing of power system data, the traditional manual data governance method can no longer meet the actual needs. The foundation of power data governance is to clarify data responsibilities and create standardized data usage standards, so as to improve the quality of organizational data, realize extensive data sharing, and apply data to strategic decision-making, resource management and business management, so as to fully display the commercial value of data assets. How to improve the data governance system, effectively standardize the workflow, manage massive data, and fully tap its value is one of the main problems that modern power industry needs to face[1-2].

Power database is usually stored and managed by structured database. There are many data formats, including text data, numerical data, logical variables and so on. Different data variables often have strong correlation or correspondence, which forms a multi-dimensional data storage scene where multiple data formats coexist. In order to ensure the harmony of data quality governance, data governance system should be regarded as a very important part of the information structure integrating data, applications, technologies and organizations [3]. How to mine the valuable information in the massive data has become a research direction that academia and industry are concerned about. Reference [4-5] analyzed the power grid big data governance system and the actual attempts of Guizhou power grid data governance, and systematically explained the concept of big data analysis and processing in the power grid, and pointed out the necessity of scientific data governance methods in the operation of the power grid. Reference [6-7] use different data governance methods to discuss in-depth grid integrated planning and power quality monitoring data. Reference [8] shows that similar association rules, clustering and outlier analysis are also used in data mining of power system operation information. Reference [9-10] describes the fast algorithm for mining association rules in large databases and the expandable framework for data cleaning, and its ideas can be used for reference in the construction of data governance methods.

However, the above-mentioned existing studies at home and abroad have not proposed a clear power data governance system. In order to further improve the data quality of the power database and more effectively improve work efficiency, this paper proposes a generalized system for power data governance, and uses the PMS to wire the data in the database. Take the original text data quality management as an example, explore the general methods of data governance rules, study the data characteristics of accurate data and inaccurate data in the PMS wire database, and use statistics and information mining methods for data quality management to assist in the establishment of data governance rules.

II. ANALYSIS OF PROBLEMS EXISTING IN POWER DATA

In the actual application of power data, the time spent on data governance is often no less than the time spent on data analysis. This is because data analysis needs to be based on high-quality data. In the case of data errors, no matter how accurate the analysis algorithm is, it will also lead to a wrong result. Therefore, a highly efficient and highly versatile data governance system is the data application The key is. In actual business processing, the original data without data preprocessing usually contains multiple types of error data.

978-1-6654-0737-3/21/$31.00 ©2021 IEEE

The categories of these erroneous data can be roughly divided into the following four categories:

(1) Data redundancy: A single record of power raw data often contains many attributes to support equipment connection management, field matching between data tables, database maintenance, etc., causing the number of attributes to exceed data analysis needs. Too many attributes of the original data are often not conducive to data modeling and analysis, so data dimension specification must be carried out.

(2) Data duplication: When the grid sensing equipment is repeatedly collected or the data storage system is self-protected and backed up, it may cause the same record to appear multiple times in the data set, although the effect of duplicate data on the data modeling results is not Large, but if the amount of repeated data is too much, it will occupy computing memory and reduce the performance of distributed computing.

(3) Data incomplete: The data acquisition equipment scattered in different voltage levels of the power grid may be due to the harsh operating environment, equipment failure, and data leakage, which may lead to the situation that the attribute value of some data is empty.

(4) Data outliers: The manufacturing process and installation defects of the equipment in each link, as well as the uncertainty in debugging and use, will bring various errors to the data, and These errors can lead to unreasonable situations in the value of an attribute in the data record. Business understanding.

III. Governance System Architecture and Analysis

Different types of erroneous data processing methods are very different. Aiming at the erroneous data existing in the original data of the power system, the data management plan formulated in this paper is shown in Fig. 1.

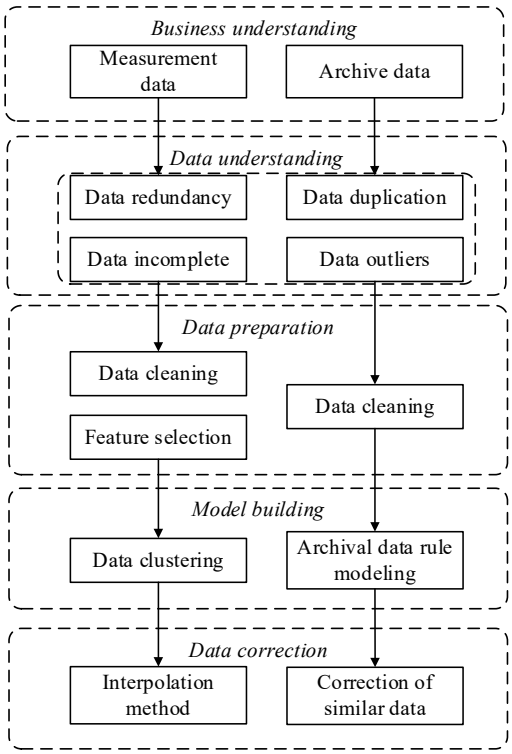

Fig. 1. Data cleaning measures

A. Business understanding

Power data governance should start from the application scenarios of power data. First, we should understand the goals and requirements of data items, and at the same time transform business knowledge into the definition of data governance issues and a preliminary plan to achieve goals.

According to scenarios, power data can be divided into . Archive data is dedicated data for power grid equipment, recording resource conditions, and supporting business management activities. Each piece of archive data is directly related to only one resource or device, and most of the field types are character types. Measurement data is the time series data that is controlled and recorded for a certain device, and most of the field types are floating point types.

B. Data understanding

Data understanding starts with the initial data collection, and the preliminary processing of the data. The purpose is to familiarize yourself with the data and identify the quality problems of the data. Because the collection of power data includes data on power operation, power supply and sales, business maintenance, etc., the main factors that affect the quality of power data are the staff's own quality and ability, post-data maintenance system, data analysis algorithm, data management mechanism, etc.

C. Data preparation

Data preparation includes all activities to construct the final data set from unprocessed data. These data will be the input values of the data governance model. Some tasks at this stage can be executed multiple times without any prescribed order. The tasks include the selection of tables, records, and attributes, as well as the conversion and cleaning of data for the data governance model.

(1) Data cleaning

Data cleaning refers to the use of data governance technologies such as mathematical statistics, information mining, and pre-defined cleaning rules to transform raw data into high-quality data that meets data quality requirements through a series of governance.

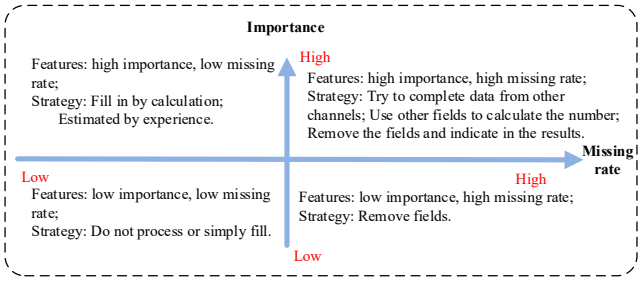

Fig. 2. Data cleaning measures

As shown in Fig. 2., according to the missing rate and importance, it is divided into removing fields, filling missing values, and re-fetching data. Among them, the methods of filling missing values are:

1) Use business knowledge or experience to speculate and fill.
2) Fill in the mean, median, quantile, mode, random value, interpolation, etc.
3) Build a model to predict the missing data.

4) Introduce dummy variables to map to high-dimensional space.

In addition, duplicate data should be removed. Duplicate data will increase the calculation time of the algorithm. In addition, the noise of the data should also be dealt with. Too much noise data will lead to poor model generalization ability. However, proper noise data can help prevent overfitting.

(2) Feature extraction and feature selection

There are many variable fields in measurement data, which require feature selection Feature selection is to select a feature subset from all the variable sets. there are three main ways of feature selection, namely filtering, encapsulation and embedded.

1) Filter type

The main idea of the filter is: "scoring" each dimension feature, that is, assigning a weight to each dimension feature, so that the weight represents the importance of the dimension feature, and then sorting according to the weight. That is, each feature is scored according to the divergence or correlation index of the feature, the score threshold or the number of thresholds to be selected are set, and the appropriate feature is selected.

2) Encapsulated type

The solution of the packaging method is not as straightforward as the filtering method. It selects an objective function to filter features step by step. According to the objective function (usually the prediction effect score), several features are selected or excluded at a time, such as recursive feature elimination algorithm (RFE). The recursive feature elimination method uses a machine learning model for multiple rounds of training. After each round of training, the corresponding features of a number of weight coefficients are eliminated, and then the next round of training is performed based on the new feature set.

3) Embedded type

The embedding method is a little more complicated. It first uses certain machine learning algorithms and models for training to obtain the weight coefficients of each feature, and selects the features according to the weight coefficients from large to small. It is similar to the filtering method, but it uses machine learning training to determine the pros and cons of features, rather than directly determining the pros and cons of features from their statistical indicators. The main idea is to learn the attributes that are best for improving the accuracy of the model when the model is established, that is, in the process of determining the model, select those attributes that are important to the training of the model.

D. Model building

At this stage, different model technologies can be selected and applied, and the model parameters are adjusted to the best values. Some technologies can solve a type of data problem that you want to pass; some technologies have special requirements in the form of data, so you need to frequently adjust back to the data preparation stage.

1) In the measurement data management, the commonly used machine learning data anomaly detection method is the clustering method. As shown in Fig. 2., clustering is mainly divided into hierarchical clustering algorithm, partition clustering algorithm, density-based clustering algorithm, grid-based clustering algorithm, model-based clustering algorithm, etc.

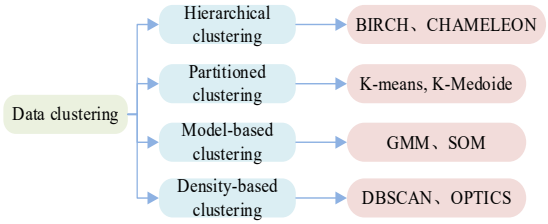

Fig. 3. Data clustering method

2) As for the model establishment of archival data, because different types of business establish different archival data modeling rules, there is usually an abnormal search for archival data based on statistics, regular expressions and feature extraction algorithms.

E. Data correction

For measurement data, there must be data omission and data integrity destruction after processing abnormal and missing values, which is not conducive to further analysis and utilization of data, so it is necessary to fill in some missing data. In order to ensure the continuity of data, interpolation algorithm is generally used to complete the measurement data. Interpolation algorithm generally adopts Lagrange interpolation algorithm, Newton interpolation method, Hermite interpolation method and so on.

For archival data, we find out the wrong data according to the established abnormal data identification rules, and then correct it according to the correct representation method of similar data.

IV. ARCHIVE DATA MANAGEMENT BASED ON TEXT PROCESSING

A. PMS business understanding

The Production Management System (PMS) is an important construction project in the informatization development of State Grid Corporation, and is one of the eight commercial applications of the SG186 project. After a long period of construction, PMS has initially realized the standardization of professional work and the informatization of management work, gradually clarifying the job responsibilities and production workflow of each post. At the same time, it also realizes the computer network management of the production business of provincial-level companies, prefecture-level companies, county-level power supply companies, and teams.

The SG186 production management system of State Grid Corporation is designed according to the design concept of five main centers: Standard Center, Equipment Center, Planning Work Center, Operation Work Center and Evaluation Center.

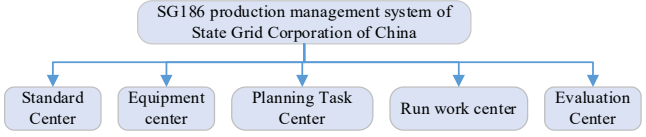

Fig. 4. State Grid PMS production management system

As shown in Fig. 4., the power grid production management system is built on the basis of the equipment center. As its core object, the equipment center plays an irreplaceable role, which is also its basic starting point and ultimate goal. The part responsible for controlling the working mode and organizing the planning task is the planning task center. This module plays a role in decision-making. In the process of implementing the process work, the operation work center is responsible for the work content and the final result of the operation. In the production management system, the evaluation supervision system and the overall value orientation cannot be handled without the grasp of the evaluation center.

B. Introduction of PMS wire data

For the PMS wire data used in this article is the basic data of the transmission part of the equipment center, in order to better express it, we select some of the parameters and related parameters needed in this article from the national grid PMS basic data entry specification (transmission part) The standardized description of is shown in Table I.

TABLE I. PMS WIRE DATA ENTRY SPECIFICATIONS (PARTIAL)

Parameter name	Unit of measure	Entry method requirements	Fill in instructions
Line name		Required, handwritten items	Subject to the naming of all levels of dispatch
Line number		Required, optional items	Subject to dispatch number
Voltage level	KV	Required, optional items	
Design voltage level	KV	Required, optional items	
Device code		Self-contained item	
Put into operation date		Required, optional items	Fill in the time when the name of the existing line was formed.
Erection method		Required, optional items	
Line length	km	Required, handwritten items	Accurate to the meter
Current properties		Required, optional items	
Owning schedule		Required, optional items	Under normal operation
Dispatch unit		Required, optional items	
Maximum allowable current	A	Required, handwritten items	Provided by the design department
Economic current	A	Optional, handwritten items	Provided by the dispatching department
Tower number		Required, handwritten items	In principle, use three-digit numbers, uniformly prefixed
Tower material		Required, optional items	
Fixed way		Required, optional items	

C. Data preparation and model building

(1) Anomaly recognition rules

Regarding the governance rules of PMS wire data, we can preliminarily establish the following content as preliminary governance rules through inquiries about the relevant regulations of the PMS business and analysis based on the characteristics of the wire data, and gradually improve these governance rules through simulation. The rules are as follows:

1) Equipment type: wire.

2) Number of assessment fields: 12.

3) Evaluation parameter items: equipment name, starting tower, ending tower, maintenance, commissioning date, equipment status, rural network, model, length (m), conductor cross section (mm2), rated ampacity (A) , Wire type.

4) Accuracy rules: 1.10≤wire cross section (mm2)≤400, 2.50≤rated current carrying capacity (A)≤800.

5) Consistency rules:

①The delivery date is earlier than the commissioning date;

② If the erection method is "mixed", any one of the overhead line length, cable line length, and total line length shall not be 0;

③ If the model includes "YJ", the overhead type is "insulated wire";

④If the model starts with "LGY", the overhead type is "bare wire";

⑤ Length (m)=(terminal tower number-starting tower number)×80;

⑥ Whether the rural network matches the regional characteristics (central area, county urban area and urban area does not match the rural network, rural area, town and township match the rural network).

6) Integrity rules: all the information in the assessment field must not be empty.

(2) Abnormal data cleaning based on regular expressions

Regular expression is a logical expression, which is usually used to deal with strings and special characters in data, also known as regular expression. By combining specific characters, we can form regular expression code matching criteria and complete the regularization of data. Regular expressions can extract and match some names, labels or letters with specific meanings in Chinese text fields, thus completing the regularization extraction of corresponding data, thus realizing the data governance function of target data. We use this method to process more than 50,000 sets of data of PMS wires, and reject the wrong data according to the anomaly recognition rules.

In the original governance rules, in the judgment of the output accuracy of the model, the model of the wire is mainly used to infer whether the wire type is correct. If the model contains "YJ", then the overhead line type should be "insulated wire", if the model contains "LGY", then the overhead line type should be "bare wire". Through the simulation of PMS data, It is found that there are different insulation and bare, rural network and non-agricultural network identifications under the same wire type, so the model and whether it is a rural network is judged according to the number of them. For example, 122>0 in the second row of Table II is judged to be an insulated wire. If 105>0 in the second row of Table III, it is judged as a rural network. As can be seen from the table, this set of regular expression search methods can find out the problematic data in the PMS wire data according to the rules, and find the loopholes in the rules to improve, and then correct the data with the wrong identification.

After simulation calculations, this paper re-established the judgment standards for the accuracy of model output and rural network output, as follows:

1) According to the majority of samples are correct and a small number of errors, find out the wrong type of wire classification.

2) According to the majority of the samples are correct and a small number of errors, find out the rural network classification errors.

TABLE II. PMS WIRE TYPE JUDGMENT (PARTIAL)

Wire type classification	Model quantity statistics [bare wire, insulated wire]	Model judgment result
Missing value	[0, 122]	insulated wire
LGJ-35	[2116, 26]	bare wire
LGJ-70	[1048, 20]	bare wire
LGJ-50/10	[534, 3]	bare wire
JKLGYJ-10-50/8	[8, 71]	insulated wire
JKLGYJ-185	[2, 3317]	insulated wire
JKLYJ-10-120	[7, 2718]	insulated wire
JKLGYJ-10-120	[3, 2639]	insulated wire
JKLGYJ-10-95/15	[1, 464]	insulated wire
LGJ-50/8	[1034, 1]	bare wire
JKLGYJ-10-185/25	[2, 508]	insulated wire

TABLE III. PMS WIRE DATA RURAL NETWORK JUDGMENT (PARTIAL)

Line classification	Statistics of rural network [rural network, non-rural network]	Rural network judgment
Jinshaqu Substation No.515 Jinlin First Circuit Line	[105, 0]	YES
Hongsibu substation 519 hongyicun line	[124, 8]	YES
Nanjiao substation 528 Kouzhuang Line	[111, 0]	YES
Pengyang substation 511 north ring road	[0, 43]	NO
Wangtuan substation 512 Railway Station Line	[162, 1]	YES
Shicaocun Line 517	[9, 58]	NO
Dawukou substation 511 Daming line	[6, 80]	NO
10kV Tongzhuang substation 511 Tongnan line	[268, 1]	YES
Hongguozi substation 525 Honghui Line	[115, 32]	YES
Honghe substation 512 Zhang He line	[149, 0]	YES

D. Abnormal data correction

By performing data cleaning on the PMS wire data, a data error statistics and index library with a huge amount of data can be obtained. Due to the huge amount of data in the established index database, there are 7722 rows of missing value data, 0 rows of wire cross-section errors, 6 rows of rated ampacity errors, 404 rows of model errors, 4302 rows of starting and ending mismatch errors, 6272 rows of length error, 1994 rows of data errors, and 476 rows of rural network classification error. We named them types 1-8 in order, and showed some indexes in the database in Table IV to analyze the characteristics of data error statistics and index libraries.

TABLE IV. ABNORMALDATA INDEX MARK (PARTIAL)

Type1	Type2	Type3	Type4	Type5	Type6	Type7	Type8
7722	0	6	404	4302	6272	1994	476
69		14900	3	23	2	55	53
82		17009	4	35	5	59	84
651		25070	5	46	56	84	140
657		37405	6	50	74	88	306
672		37406	7	53	80	104	320
673		48416	8	63	91	125	330
679			9	92	96	126	332
680			10	114	97	141	335
681			11	119	134	142	357

The first row of the abnormal data index indicates the judgment rule corresponding to the column, and the second row indicates the number of errors of this type. The numerical value corresponding to each row under this column represents the number of rows of data with problems in the original PMS database. There are 7722 missing values in PMS database, which indicates that many missing data in PMS data reflect the incompleteness of PMS data. There are relatively few errors in the database of conductor cross section and rated current carrying capacity, only 0 and 6 respectively, but as important parameters of PMS conductor, it is still necessary to control them. The analysis results show that there are a large number of errors in conductor type, matching between start and end, conductor length and date, and the judgment of rural power grid. Therefore, Table IV confirms that there are data missing, data mismatch, data error and poor accuracy in PMS conductors, which shows that it is necessary to control the data quality of PMS conductors.

V. CONCLUSION

In order to improve the reliability and scientificalness of power massive data analysis, so that power grid companies can use and mine a wide range of different types of data more effectively, this paper puts forward a set of processes suitable for power grid data management. High-quality data can be obtained through fine analysis of five steps: business understanding, data understanding, data preparation, model building and data correction. Combined with the analysis and simulation of PMS archives data management process, the practicability and effectiveness of the proposed framework are verified.

REFERENCES

[1] C. Ming, X. Yanbin, W. Shengyan, et al. Data Assets Governance and Related Key Issues for Energy Internet Based on Data Curation Theory [J]. Power Grid Technology, 2020,44(7):2420-2429.

[2] M. Janssen, P. Brous, E. Estevez, et al. Data governance: Organizing data for trustworthy Artificial Intelligence[J],Government Information Quarterly,2020,37(3):1-8.

[3] J. Kezhen, W. Zhenzhen. Research on Power Enterprise Data Governance System [J]. Power Information and Communication Technology, 2014,12(1):7-11.

[4] T. Yun. A Preliminary Study on Big Data Governance System of Power Grid [J]. Electronic Technology and Software Engineering, 2017, 1(5): 182-183.

[5] J. Yuan. Guizhou Power Grid Data Governance in Big Data Era [J]. Power Big Data, 2017, 20(8): 88-92.

[6] Z. Yi, Y. Honggeng, Y. Maoqing. Management scheme of massive power quality monitoring data based on distributed file system [J]. Power System Automation, 2014, 1(2): 108-114.

[7] L. Guojiao. Application Research of Power Quality Analysis Based on Big Data [D], Master's Degree Thesis, Changchun: Changchun University of Technology, 2019.

[8] L. Zhiyong. Research on Data Mining of Power System Operation Information [D], Master's Degree Thesis, Hangzhou: Zhejiang University, 2009.

[9] G. Yangyang, X. Liewei, Y. Jian. Design of a new dynamic reconfigurable regular expression matching engine [J]. Journal of Fudan University (Natural Science Edition), 2019, 58(6):706-718.

[10] H. Xiaohui. Design and Implementation of Training Simulation Platform for Decentralized and Autonomous Dispatching Centralized System (CTC) [D], Master's Degree Thesis, Shanghai, East China Jiaotong University, 2019.