Advances in Artificial Intelligence, Big Data and Algorithms
G. Grigoras and P. Lorenz (Eds.)

doi:10.3233/FAIA230902

A Study of Data Governance Methods Based on Non-Linear Evaluation of K-Domains

Yajie LI[a,b1], Kaining SUN[b,c], Tao MING[a,b], Anqing CHEN[d]
[a]*State Grid Xinjiang Electric Power Co., Ltd. information and communication company, Urumqi, Xinjiang, China 832000*
[b]*Xinjiang energy Internet big data laboratory, Urumqi, Xinjiang, China 832000*
[c]*State Grid Xinjiang Electric Power Co., Ltd, Urumqi, Xinjiang, China 832000*
[d]*State Grid Xintong Yili Technology Co., Ltd., Fuzhou, Fujian, China 350000*

Abstract: The current traditional data governance methods mainly ensure the overall quality of data through data quality inspection, and the lack of data fusion processing leads to poor data governance results. In this regard, a data governance method based on nonlinear evaluation of K domain is proposed. By formulating data cleaning rules, data are processed such as weight reduction, gap filling and deletion, and data fusion operations are performed according to the fusion table attributes by configuring fusion rules. Finally, the data consistency as well as wholeness is checked, and the actual effect of data governance is determined by the data quality score results. In the experiments, the proposed method is validated for the governance effect. The analysis of the experimental results shows that the proposed method takes significantly shorter time for data fusion and has better data governance performance when the proposed method is used to govern multiple sources of data.

Keywords: k-domain; nonlinear assessment; data governance; data quality dimensions

1. Introduction

At present, the governance of multi-source data can be carried out in two main ways, namely, building a data quality management model and building a source data quality framework. Before building the data quality management model, the original data needs to be iterated according to the governance requirements, and the data will have the governable value only when the number of iterations of the original data reaches the standard [1]. By extracting and analyzing different quality attributes of the data, the most relevant attribute characteristics of the core data are extracted, and the data quality management model built on this basis can effectively represent the characteristics of the original data, so that the data can be efficiently governed. Similar to the data quality management model, the data quality framework also requires pre-processing of the original data to improve data purity and reduce data redundancy. The data quality framework is built by analyzing the syntax and semantics of the data, extracting the

[1] Corresponding author: Yajie LI, State Grid Xinjiang Electric Power Co., Ltd. information and communication company; Xinjiang energy Internet big data laboratory; e-mail: jie.15@163.com

core data semantics, and combining this with computer programs to build a hierarchical framework related to data quality. In this framework, the input data are subject to a strict set of specifications, from data generation to data processing results output, in terms of format and content. Therefore, the data generated on the data quality hierarchy framework can also achieve the purpose of improving the overall data quality and achieving data governance. Both of these conventional data governance methods can standardize and process data to a certain extent to obtain reliable data with higher accuracy and consistency, and are widely used in many business areas. However, these two conventional data governance methods also have certain limitations, which are reflected in the scale of data and the poor data processing effect. First, in terms of data scale, the conventional data governance methods treat all data as two-dimensional data due to the lack of effective analysis of data dimensionality, which leads to some multidimensional data being severely compressed. On this basis, the completed data after processing loses its multidimensional characteristics, thus making the completed data differ greatly from the original data, and its data attributes are missing too much to replace the original data for processing [2]. In this regard, the data dimensions can be classified according to the different data attributes of the actual data as well as the data types, and according to the characteristics of the data attributes. For data with a relatively single data composition structure, the data can be processed according to one-dimensional or two-dimensional data. For multi-field data with more complex attribute composition, the dimensions of the data can be divided according to the number or characteristics of the attributes, so as to quantify the data dimensions and thus avoid the situation of missing data dimensions. The data after processing on this basis is not only consistent with the source data, but also the data dimensions are not destroyed and can represent the original data for the next operation. In terms of data processing effect, the conventional data governance methods cannot determine the effectiveness of the data processing method due to the lack of scoring of data quality, which leads to the inability to determine the effectiveness of the data processing effect after the data processing is completed. In this regard, this paper combines the K-domain nonlinear evaluation theory and proposes a scientific data quality scoring means for the specific steps of data processing to score the data governance effect, so as to effectively and efficiently judge the practicality of data governance methods [3].

2. Data Cleaning

Data cleansing is the first step in data governance. By cleaning the data, the lower quality data or duplicate data can be effectively filtered out to improve the quality of data and thus achieve flexible use of data [4]. Data cleaning should be carried out in accordance with the established cleaning principles.

Four data cleaning processes are performed on the raw data, namely data deletion, data retention, data de-duplication and data formatting. Among them, data deletion includes deletion of rows and columns, as well as deletion of spaces and duplicate fields [5]. Data retention includes the retention of numbers and letters, and the replacement of long fields. Data de-duplication refers to the processing of two sets of data with the same data source, by adding a prefix or suffix to one set of data to distinguish it from the other, thus achieving the effect of data de-duplication. For duplicate data with prefix or suffix, and the content of the data is less related to the core

database, the data can be directly eliminated in the way of data weight reduction processing. The specific data cleaning process is shown in the Figure 1.

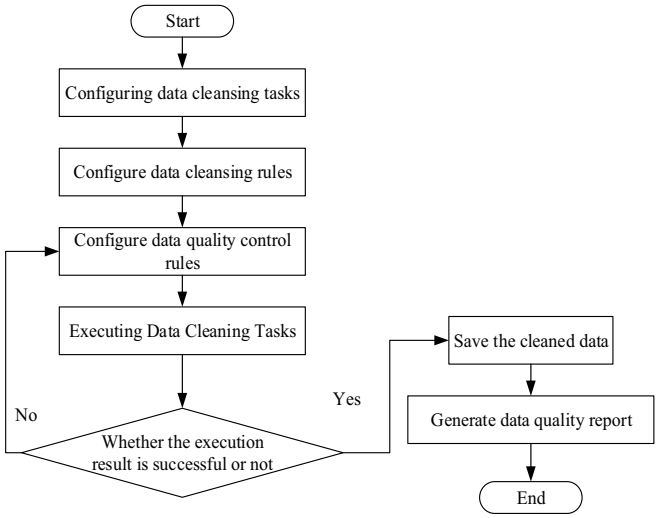

Figure 1. Data cleaning process

According to the above figure, before data cleaning, users need to first configure the cleaning task, which includes configuring the cleaning rules, cleaning data names and data quality control rules. After the configuration of the data cleaning rules is completed, the data cleaning task can be started at [6]. After the task is executed, the algorithm will determine the result of the cleaning task, and if the cleaning task is successful, the cleaned data will be stored and a report about the data quality will be generated. If the data cleaning task cannot be executed normally according to the process, it returns to the data cleaning task configuration and checks the data source name, data source and data cleaning rules and other related parameters configured to determine whether the configuration content and the corresponding parameters meet the set standards, and then re-executes the data cleaning task after the check is correct. The spark algorithm is chosen as the main execution algorithm for data cleaning. This algorithm can complete large-scale data cleaning in a relatively short time, with high data processing performance and low impact on the load of the data system. Users can configure the data cleaning parameters using the spark algorithm according to the actual data calling requirements. The algorithm automatically generates two types of data, namely field data and non-field data, based on the actual content of the data package, and cleans the data according to the different data types. In addition, as the process of cleaning data, there are several cleaning rules that can be followed. Different data cleaning rules correspond to different data governance needs, so before data cleaning, multiple data cleaning rules can be configured according to the actual needs of data governance. spark algorithm will transform the data cleaning rules, so that they can de-weight and replace redundant data and low-quality data, thus improving data purity [7].

The above steps will complete the effective cleaning of data, store the cleaned data into the database, and use the operator to record the cleaning results and the cleaning process.

3. Multi-Source Data Fusion

The source data for this data governance includes three main parts, namely business data, technical data and process data. Among them, business data includes data related to business content, technical data includes storage information and field information, and process data includes data change information and work output information.[8]. In order to achieve effective governance of the overall data, the three types of data need to be fused with multiple sources of data to improve the correlation between the data and to improve the efficiency of data governance. Since different data types are stored in different locations and there are differences in the data acquisition methods, before fusing the multi-source data, it is necessary to obtain the data information of different data sources, and the specific data types and the corresponding acquisition methods are shown in the Table 1.

Table 1. Data information acquisition methods

Number	Data Type	Data Storage Location	Source Data Information	Data Acquisition Method	Data Information Update
1	MySQL	information_schema	(TABLES/COLUMNS/VIEWS/PARTITIONS)	JDBC	Regularly query the most recently changed data tables and data table structures from the management system
2	Hiive	hive	File Storage Information, Data property information	JDBC	Scan queries every half minute to get the most recent data content, then work through Hive JDBC for table structure synchronization
		Default Storage Location	Field Information, Partition Information		

The above table shows that the data types mainly include MySQL data and Hive data, which correspond to the basic data information and file storage information respectively. In order to improve the efficiency of data acquisition, we choose the pull method to call the data, and update the data table structure according to the regularly updated data information. After obtaining data from different data sources, data fusion can be carried out. There are two main types of data fusion, namely, single-base fusion and multi-base fusion. Among them, single-base fusion does not involve database selection, and data fusion can be performed directly by configuring fusion rules and according to the fusion table properties. Multi-base fusion, on the other hand, involves data from different databases, and the details such as structure and fields of the data are different from each other. Therefore, before performing multi-database fusion, you need to select the type and name of the database, categorize the data with the same or similar data structure, and then perform the corresponding fusion operation according to the fusion rules. The data fusion rules are randomly generated according to the content of the selected data table, and each rule can represent the content attributes of the data table. The fused data properties are selected according to the fusion requirements, and the data tables are named, and the fused data is saved in the new

named tables. Once the fusion rules are configured through the above steps, the fusion operation can be performed according to the corresponding fusion rules, and the fused data will be converted into SQL statements for database reading and storage. The fused data is stored in the database as a new table to distinguish it from the un-fused data. The user can also change the form of the fusion result presentation and store the fused data according to their needs, or save the fused data as other data types if they want to use the fused data for analysis in other software.

The above steps can complete the processing and fusion of data from multiple sources, providing a more consistent data source for subsequent data governance.

4. Data Quality Audits Based on Non-Linear Assessment of K Domains

Data quality audit is the core work of data governance, through the inspection of data quality, the problematic data that adversely affect the overall data quality can be extracted, so as to ensure the rationality of the data. In this paper, the audit of data quality needs to be carried out in accordance with the specified audit rules, and the specific data audit rules include missing data check, data integrity check, data logic check, data format specification check and data real-time check.

According to the above process, the audit tasks need to be configured before the data quality is audited. Among them, the audit task mainly involves the content to be audited, the description of the audit password, the audit object and the number of audits. According to the actual needs of the audit task configuration is completed, you can carry out data auditing work. The task of data auditing is to ensure the integrity of the data while effectively checking the consistency of the data, so as to provide assurance of data quality. Therefore, it is necessary to configure the corresponding data inspection rules before conducting the audit. The data inspection rules include checking the attributes and fields of the data to see if there are blank data fields. In addition, the internal structure of the data and the degree of association between the data should be checked to determine whether the internal structure of the data and the data connectivity meet the internal logic of the database. At the same time, in order to ensure the uniqueness of the data, the primary key of the data should be reduced so that each data has a unique primary key to correspond to it. The data real-time inspection refers to the inspection of the number of data updates to ensure that the frequency of data updates is consistent with the actual frequency, so as to ensure that the real-time data will not be affected. After completing the data audit, the algorithm will determine the data audit results, if the data audit task is successfully executed, the corresponding audit records and data quality reports can be generated; if the audit task is not effectively carried out, then return to the audit task configuration step, check the parameters and details of the audit task configuration, and continue the data audit work after the reconfiguration is completed.

After completing the data audit task, in order to make the data audit results more adequate and more intuitive to display in front of users, it is also necessary to score the data audit results. The data quality scoring incorporates the K-domain nonlinear evaluation theory and searches the scoring results based on the K-domain to get the optimal scoring results, and the specific scoring formula is shown in Eq. (1).

$$S = \frac{S_{qualified}}{S_{total}} \times 100\% \tag{1}$$

Where, S represents the data scoring result after the audit, $S_{qualified}$ represents the number of data that meet the data audit rules, and S_{total} represents the total amount of data in the source database. By completing the scoring of data through the above formula, the quality of data audit can be effectively grasped. Combining the contents of this section with the above-mentioned contents related to data cleaning and multi-source data fusion, the design of data governance method based on non-linear evaluation of K domain is completed at.

5. Testing and Analysis

5.1. Test Preparation

To prove that the data governance method based on nonlinear evaluation of K domain proposed in this paper outperforms the conventional data governance method in terms of governance effect, an experimental test session is constructed to verify the actual governance effect of this data governance method after the theoretical part of the design is completed. To ensure the reliability of the data results, this experiment is conducted in a comparative experiment, and two conventional data governance methods are selected as the experimental control group, namely, the lifecycle-based data governance method and the Linked Data-based data governance method.

The data set used for this test is derived from both MySQL as well as Hive data, and the data table types are all test data. The MATLAB software was used to build the simulation data model, and the data in the dataset required for the experiment were uploaded to the test model. Three data governance methods are used to simulate the test data in the data source and compare the parsing results obtained by the software, and the specific experimental data set information is shown in the following Table 2.

Table 2. Experimental data set information

Experimental test data information	Data description
Test data preparation	Data Source Type: MySQL/Hive Data source name: Test Data Data Table Name: gd_test basic_info gd_result gd_service Data Volume: 5000 items selected according to the number of experiments
Test Objectives	Does the length of time spent on the multi-source data fusion feature for MySQL and Hive data types meet business requirements
Test Methodology	Perform data fusion operations on MySQL and Hive data sources respectively; 2. Select the specified three tables, select the attributes, retain 42 attribute columns, and set the condition to id equal; 3. The rest of the page in accordance with the prompts to complete the configuration of parameters in turn, click "Finish Fusion" operation
Completion criteria	Each application can respond within the specified time and the data is loaded without loss

The above experimental parameters were used to configure the experimental environment. To improve the comparability of the experimental results, this experiment was conducted five times, with increasing amounts of test data, to determine in turn the time spent on data fusion by data governance methods for different amounts of data.

5.2. Analysis of Test Results

The evaluation index chosen in this paper is the performance of the data governance method, which is measured by the time spent by the data governance method in fusing data from multiple sources, and the less time spent means the actual performance of the data governance method is better. In this experiment, two different data sources are selected to test the method, and the specific experimental results are shown in the Figure 2.

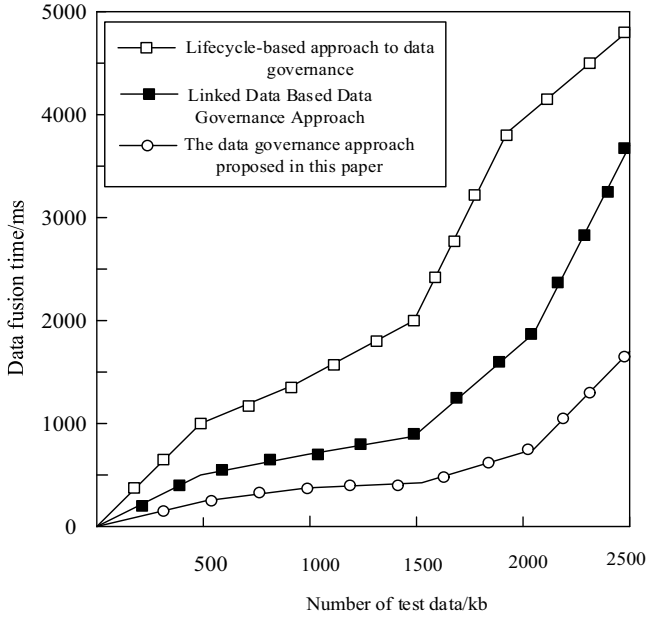

Figure 2. Comparison of data fusion time

Based on the above experimental results, it can be seen that there are also differences in the data fusion time under different data governance methods when data fusion processing is performed on different data sources. The numerical comparison clearly shows that the two conventional data governance methods take longer time to fuse the data, and the longest fusion time can reach 5000ms, while the data governance method based on K-domain nonlinear evaluation proposed in this paper takes significantly shorter time to fuse the data from different data sources, and the fusion time does not fluctuate with the growth of the tested data volume. The data fusion time is significantly shorter when fusing data from different data sources, and the fusion time does not fluctuate significantly with the growth of test data volume. It can be seen that the data governance method proposed in this paper has better data governance performance and can effectively govern large-scale data.

6. Conclusion

In this paper, the data governance method based on the nonlinear evaluation of K domain is proposed. By formulating data cleaning rules, a total of four data cleaning processes are performed on the original data, namely, data deletion, data retention, data de-duplication and data format adjustment. And the data with the same or similar data structure are categorized, the data are processed for multi-bank fusion, and finally the data quality is checked, which mainly includes data consistency check and data integrity check. It is conducive to improving data processing efficiency and meeting the needs of large-scale data governance.

References

[1] Ahmadi S, Tavana M M, Shokouhyar S, et al. A new fuzzy approach for managing data governance implementation relevant activities[J]. The TQM Journal, 2022, 34(5): 979-1012.
[2] Ji X , Niu P , Wang P . Non-existence results for cooperative semi-linear fractional system via direct method of moving spheres[J]. Communications on Pure & Applied Analysis, 2020, 19(2):1111-1128.
[3] Poddar A , Kumar N , Kumar R , et al. Evaluation of non-linear root water uptake model under different agro-climates [J]. Current Science, 2020, 119(3):485-496.
[4] Son V V , Duong D T , Hoang T M , et al. Analysing outage probability of linear and non-linear RF energy harvesting of cooperative communication networks [J]. IET Signal Processing, 2020, 14(8):541-550.
[5] Khurana N . Issue Analysis: A Use-Driven Approach to Data Governance Can Promote the Quality of Routine Health Data in India[J]. Global Health: Science and Practice, 2021, 9(2):238-245.
[6] Al-Dossari H , Sumaili A A . A Data Governance Maturity Assessment: A Case Study of Saudi Arabia[J]. International Journal of Managing Public Sector Information and Communication Technologies, 2021, 12(2):19-30.
[7] Liu X , Sun S X , Huang G . Decentralized Services Computing Paradigm for Blockchain-Based Data Governance: Programmability, Interoperability, and Intelligence[J]. IEEE Transactions on Services Computing, 2020, 13(2):343-355.
[8] Yang Q , Wang H , Li H , el al. On Line Identification of Quantized Stochastic Nonlinear Systems Based on Quadratic Programming [J]. Computer Simulation,2022(7) 324-330.