A Data Quality Improvement Method Based on the Greedy Algorithm

Zhongfeng Wang[1,2], Yatong Fu[1,2], Chunhe Song[1,2(✉)], Weichun Ge[3], Lin Qiao[3], and Hongyu Zhang[3]

[1] Key Laboratory of Networked Control Systems, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, People's Republic of China
songchunhe@sia.cn
[2] Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110016, China
[3] State Grid Liaoning Electric Power Co., Ltd., Shenyang 110000, People's Republic of China

Abstract. High-quality data is very important for data analysis and mining. Data quality can be indicated by many indicators, and some methods have been proposed for data quality improvement by improving one or more data quality indicators. However, there is few work to discuss the impact of the processing order of data quality indicators on the overall data quality. In this paper, first, some data quality indicators and their improvement methods are given; second, the impact of the processing order of data quality indicators on the overall data quality is discussed, and then a novel data quality improvement method based on the greedy algorithm is proposed. Experiments have been shown that the proposed method can improves the data quality while reducing the time and computational costs.

Keywords: Data quality · Improvement order · Greedy algorithm

1 Introduction

With the rapid development of Internet and information technology, data has become an important asset and competitive resource of enterprises. Almost all industries can benefit from data, but the problems of missing key fields, too much data noise and confused data classification in the original data make the quality of the original data too low. Low-quality data will directly affect the value of data analysis and mining. Data quality problems may occur in all stages of data acquisition and storage. How to effectively improve data quality has always been a difficult problem to solve. There are three main reasons. One is the lack of a standard definition of data quality. All walks of life have different understandings and needs for data quality. It is difficult to form a standardized and unified data quality standard. Secondly, it is difficult to determine data quality indicators. Data quality indicators are the process of quantifying data quality. Now the research on data quality indicators only focuses on a specific field, and there is no universal evaluation framework. Thirdly, data quality indicators are not independent,

X. B. Zhai et al. (Eds.): MLICOM 2019, LNICST 294, pp. 256–266, 2019.
https://doi.org/10.1007/978-3-030-32388-2_22

and the improvement of one data quality indicator may result in the reduction of another data quality indicator. How to determine an efficient data quality improvement strategy is a problem to be studied at present.

Data quality has already attracted extensive attention of researchers, and the research on data quality indicators is becoming more and more mature. However, the research on establishing efficient and reasonable data processing order based on data quality indicators is at an early stage. At present, the research on data quality indicators mainly focuses on the correlation between different indicators, but the correlation between indicators is the follow-up. However, it is difficult to prove the benefits of data processing. In determining the order of data processing, the current research mostly adopts traversal enumeration, which is a relatively inefficient method. In view of the current research situation, this paper first summarizes the commonly used indicators for data quality quantification and the methods used to improve each indicator; secondly, it analyses the impact of data quality improvement execution order on the overall data quality; finally, a greedy algorithm for data quality improvement is proposed, and an efficient and reasonable data processing order is determined. The validity of the data processing order is proved by simulation experiments.

The rest of this paper is organized as follows. Section 2 outlines the related works. Section 3 introduces data quality and data quality improvement methods. Section 4 analyzes the impact of the order on data quality improvement. Section 5 shows experimental results and analysis. Section 6 concludes this paper.

2 Related Works

In the field of data quality, Cai et al. have made a very detailed study of the history of data quality [1]. Researchers began to study data quality in the 1970s. At that time, although there was no knowledge system of data quality, they have found that poor quality data will have a negative impact on information systems. Saha and Srivastava point out that poor quality data are common in large databases and networks [2]. Poor quality data will have a serious impact on the results of data analysis. Data quality research was formally carried out in the 1990s, and data quality definition and measurement indicators began to be established. At present, data quality research is booming, data processing algorithms and frameworks continue to emerge, international organizations began to research and develop data quality standards. In terms of data quality indicators, [2–4] have conducted in-depth research on data quality dimensions. Wang et al. carried out extensive research work, investigated 118 kinds of data quality properties, summarized 20 kinds of commonly used data quality properties [2]. Sidi et al. analyzed 40 kinds of data quality properties in detail [3]. Zaveri et al. unified common terms of data quality, provided 18 data quality dimensions and 69 kinds of data quality measurement methods [4]. Wang et al. expounded the connotation of scientific data and data quality, studied the basic principles of data quality evaluation, analyzed the data quality structure, and put forward the scientific data quality evaluation index system based on basic level, criterion level and index level [5]. In [7], the problem of determining the timeliness of a set containing redundant records under a given time limit is studied, and an algorithm for solving the problem of determining the

timeliness is proposed for the first time. [8] studies data integrity in detail. In data restoration, [9] studies record matching and data restoration, and points out that data quality indicators are not independent of each other. In [10], Fan et al. studied the new problems related to data cleaning, namely the interaction between record matching and data repair, and proved that data repair can effectively help identify matches and improve data quality. In [11], Gackowski et al. conducted a preliminary study on the establishment and validation of data quality indicators, and explored the logical relationship between data dimensions. [12–14] explores the correlation between different data quality indicators. Ding et al. proposed precise definitions and violation patterns of four data quality indicators, such as timeliness and accuracy [12]. Cheng et al. discussed the impact of the relationship between different data quality indicators, and determined an effective data cleaning strategy for data in wireless sensor networks [13]. Dominikus proposed improving data quality operations for multi-dimensional data quality assessment [14]. In [15], Helfert analyzed the dependence of data quality dimension, studied how to evaluate the overall quality of data by the total weighted dimension score, and examined the applicability of this method. [16] provides a wide range of technologies for evaluating and improving data quality, focusing on data quality assessment and improvement techniques. Zhao et al. comprehensively analyzed the content and method of quality evaluation of correlated data, providing reference for the construction of quality control and evaluation system [17]. In [18], Liu briefly analyzed the causes of some statistical dishonesty problems described in this paper, and proposed measures to improve the quality of statistical data. [19] uses association rules to evaluate data quality. Reza et al. found that most of the indicators developed so far do not take data weight into account, thus defining a new measurement standard based on data weight to further improve its data quality [20].

Table 1. Nomenclature

d_{ijt}	The j-th dimensional of the data d sampled from the i-th data source at the time t, where $i \in N$ and $j \in M$, and N and M are the numbers of data sources and the number of dimension of d
d_{ij}	The j-th dimensional of the data d sampled from the i-th data source over all time
d_{ijq1}, d_{ijq2}	Lower quantile and upper quartile of d_{ij}
$d_{ij\Delta q}$	Difference between d_{ijq1} and d_{ijq2}
Δt	Sampling interval
Δt_{mean}	Mean sampling interval
S	All data in the database

3 Data Quality and Data Quality Improvement

In this section, first, the precise definition of data quality model is given. Second, the methods used to improve each index are summarized. Finally, the impact of data quality improvement execution order on the overall data quality is analyzed, which lays

the foundation for the next section of experimental simulation. For convenience, some nomenclatures used in this paper are given in Table 1.

3.1 Data Quality Indicators

Currently there are many data quality indicators, and in this paper, some most popular data quality indicators, including data integrity, data accuracy, data consistency, and data timeliness are used and analyzed.

Data integrity indicator $D_{\text{integrity}}$: is used to measure data integrity, including scale integrity, attribute integrity, content integrity and so on. Data integrity can be measured by data size, data volume, data coverage and so on. In order to simplify the data quality assessment model, in this paper, data integrity is defined as the degree of data missing. The following is the accurate definition of data integrity:

$$s_{ijt} = \begin{cases} 1, d_{ijt} \neq None \\ 0, d_{ijt} = None \end{cases} \tag{1}$$

$$D_{\text{integrity}} = \frac{\sum_{t\in T}\sum_{j\in M}\sum_{i\in N} s_{ijt} - S}{\sum_{t\in T}\sum_{j\in M}\sum_{i\in N} s_{ijt}} \times 100\% \tag{2}$$

Data accuracy indicator D_{accuracy}: is to measure the ability of data to accurately describe the physical world. Illegal values, invalid data types and low data accuracy can all be used to measure data accuracy. In order to simplify the data quality evaluation model, this paper defines data accuracy as the degree of data anomalies. The following is the precise definition of data accuracy:

$$s^a_{ijt} = \begin{cases} 1, d_{ijt} \notin \left[d_{jq1} - 1.5 \times d_{j\Delta q}, d_{jq2} + 1.5 \times d_{j\Delta q}\right] \\ 0, d_{ijt} \in \left[d_{jq1} - 1.5 \times d_{j\Delta q}, d_{jq2} + 1.5 \times d_{j\Delta q}\right] \end{cases} \tag{3}$$

$$D_{\text{accuracy}} = \frac{\sum_{t\in T}\sum_{j\in M}\sum_{i\in N} s_{ijt} - \sum_{t\in T}\sum_{j\in M}\sum_{i\in N} s^a_{ijt}}{\sum_{t\in T}\sum_{j\in M}\sum_{i\in N} s_{ijt}} \times 100\% \tag{4}$$

Data consistency indicator $D_{\text{consistency}}$: is to measure the consistency of different data formats, contents and ranges of single or multiple data sources. It has been pointed out in [7] that data consistency includes conceptual consistency, format consistency, range consistency and time consistency. Different constraints are determined according to the data content. Data that violate the constraints are considered to have a consistency conflict. The more complex the constraints are, the more complex the consistency discrimination is. In order to simplify the data quality evaluation model, this paper only establishes a constraint condition: the sampled value range of the same node on the

same attribute should be consistent. In this paper, data consistency is defined as the extent to which data violates constraints, using Eqs. (5) and (6):

$$s_{ijt}^{c} = \begin{cases} 1, d_{ijt} \notin \left[d_{ijq1} - 1.5 \times d_{ij\Delta q}, d_{ijq2} + 1.5 \times d_{ij\Delta q}\right] \\ 0, d_{ijt} \in \left[d_{ijq1} - 1.5 \times d_{ij\Delta q}, d_{ijq2} + 1.5 \times d_{ij\Delta q}\right] \end{cases} \tag{5}$$

$$D_{\text{consistency}} = \frac{\sum\limits_{t\in T}\sum\limits_{j\in M}\sum\limits_{i\in N} s_{ijt} - \sum\limits_{t\in T}\sum\limits_{j\in M}\sum\limits_{i\in N} s_{ijt}^{c}}{\sum\limits_{t\in T}\sum\limits_{j\in M}\sum\limits_{i\in N} s_{ijt}} \times 100\% \tag{6}$$

Data timeliness indicator $D_{\text{timeliness}}$: is to measure the freshness and availability of data. It refers to the time interval and efficiency of receiving, processing, transmitting and utilizing information from the information source. The shorter the time interval, the more timely information updates, the more time-sensitive data. In order to simplify the data quality assessment model, in this paper, data timeliness is defined as the degree of data update, using Eqs. (7) and (8):

$$s_{it}^{t} = \begin{cases} 1, \Delta t > 2 \times \Delta t_{mean} \\ 0, \Delta t \leq 2 \times \Delta t_{mean} \end{cases} \tag{7}$$

$$D_{\text{timeliness}} = \frac{\sum\limits_{i\in N}\sum\limits_{t\in T} s_{it} - \sum\limits_{i\in N}\sum\limits_{t\in T} s_{it}^{t}}{\sum\limits_{t\in T}\sum\limits_{i\in N} s_{it}} \times 100\% \tag{8}$$

According to the definition of data quality index above, the total quality of data is recorded as Q, the measure value of data quality index x is recorded as Q, and the weight of data quality index x is recorded as Q. Then the precise definition of data quality is defined as:

$$Q = \sum_{x\in X} \omega_x D_x \tag{9}$$

3.2 Data Quality Improvement Methods

According to the definitions of data integrity, data accuracy, data consistency, and data timeliness, there are many methods can be used to improve the data quality. In order to make the scheme proposed in this paper clear and comparable, following data quality indicators improvement methods are used:

The operation of improving data integrity $P_{\text{integrity}}$: for improving data integrity operation, there are mean interpolation, similar mean interpolation, modeling prediction, high-dimensional mapping, multiple interpolation and other methods. In this paper, the method of improving data integrity based on mean interpolation is used.

The operation of improving the accuracy of data $P_{accuracy}$: the abnormal data in data sets is the main reason for the low accuracy of data. The abnormal data can be

identified by using data statistics technology and data visualization. The commonly used methods are box-dividing method, regression method, clustering method and so on. In order to improve the accuracy of data, this paper identifies the abnormal data and fills it with the mean value.

The operation of improving data consistency $P_{consistency}$: when data violates constraints, it is regarded as abnormal data, and the abnormal data is filled up according to constraints rules to improve data consistency.

The operation of improving data timeliness $P_{timeliness}$: judging the time items in the data set and deleting or filling the data with low timeliness.

4 Data Quality Improvement Based on the Greedy Algorithm

4.1 Analysis of the Impact of the Order on Data Quality Improvement

Before data processing, it is necessary to clarify the impact of the order of data quality improvement on the overall quality of data. Specifically, the following points need to be noted:

First, the improvement of a certain data quality indicator cannot ensure the improvement of the overall data quality. For example, suppose data consistency is denoted by $D_{consistency}$, and the total data quality is denoted by Q. After the data consistency operations, data consistency is $D'_{consistency}$ and the overall data quality is Q'. Although $D_{consistency} < D'_{consistency}$, Q' may be less than Q, that is to say, increasing individual data quality indicators does not always improve the overall data quality.

Second, individual data quality gains cannot be directly accumulated. For example, data integrity is $D_{\text{integrity}}$, data consistency is $D_{consistency}$, and the total data quality is denoted by Q. For a dataset, when carrying out the data integrity improvement operation $P_{\text{integrity}}$ individually, the resulted data quality is Q', and the data quality gains is $\Delta Q' = Q' - Q$; when carrying out the data integrity improvement operation $P_{consistency}$ individually, the resulted data quality is Q'', and the data quality gains is $\Delta Q'' = Q'' - Q$; when carrying out $P_{\text{intergrity}}$ and $P_{consistency}$ sequentially, the resulted data quality is Q''', and the data quality gains is $\Delta Q''' = Q''' - Q$; at this time, $\Delta Q''' \neq \Delta Q' + \Delta Q''$. In other words, data quality gains cannot be directly accumulated.

Third, different processing order of individual data quality indicators results in different overall data quality gains. For example, when carrying out $P_{\text{integrity}}$ and $P_{consistency}$ sequentially, the resulted data quality is Q, and the data quality gains is $\Delta Q' = Q' - Q$; when carrying out $P_{consistency}$ and $P_{\text{integrity}}$ sequentially, the resulted data quality is Q, and the data quality gains is $\Delta Q'' = Q'' - Q$; at this time, $\Delta Q' \neq \Delta Q''$. That is to say, different data processing order and different data quality gains.

Fourth, more data quality improvement operations cannot ensure better overall data quality gain. For example, data integrity is $D_{\text{integrity}}$, data consistency is $D_{consistency}$, and the total data quality is denoted by Q. For a dataset, when carrying out the data integrity improvement operation $P_{\text{integrity}}$ individually, the resulted data quality is Q', and the data quality gains is $\Delta Q' = Q' - Q$; when carrying out $P_{\text{integrity}}$ and $P_{consistency}$ sequentially, the resulted data quality is Q'', and the data quality gains is $\Delta Q'' = Q'' - Q$; at this

time, $\Delta Q''$ is not always greater than Q'. That is to say, more data quality improvement operations cannot ensure better overall data quality gain.

4.2 The Proposed Method

Assuming that each data quality indicator corresponds to a data processing operation, and that each data processing operation is not reused. Suppose there are n data processing operations, then the total number of possible orders is:

$$O_n = \sum_{j=0}^{n} \left(\prod_{i=0}^{j} (n-i) \right) \tag{10}$$

For example, when there are four data quality indicators, there will be 64 data processing orders to choose. When there are five data quality indicators, there will be 325 kinds of data. When the dimension of data quality index increases, the situation will become more complex. To find the optimal data processing strategy, it is necessary to traverse all data processing strategies. In order to save time and cost, it is a reasonable way to use greedy algorithm to determine the order of data processing.

The flowchart of the proposed method is shown in Fig. 1.

Suppose there are n data processing operations in the data operations set DO.

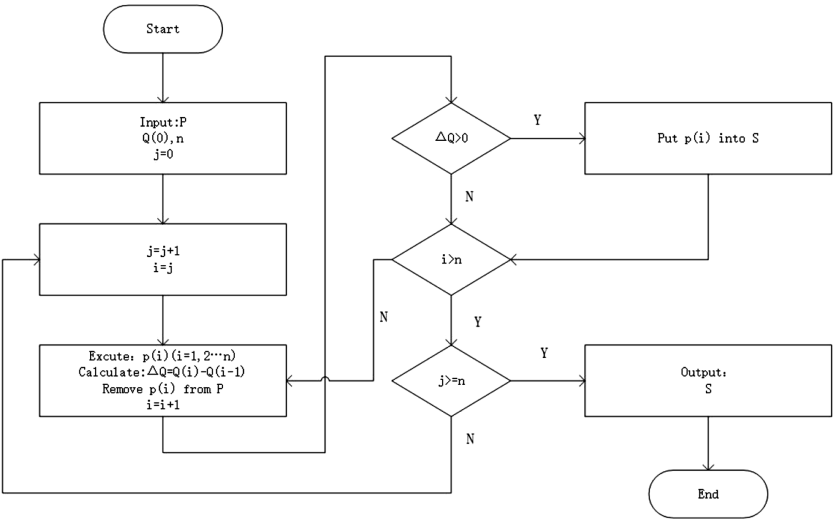

Fig. 1. The flowchart of the proposed method.

Step 1: select the i-th data processing operation P_i in DO, carry out data indicator improvement using P_i, suppose the resulted overall data quality is Q_i.
Step 2: calculate the data quality gain $\Delta Q = Q_i - Q_{i-1}$. If $\Delta Q > 0$, then put P_i into the operations set S_j . $i = i+1$, go to Step 1, until $i = n$.
Step 3: $j = j+1$, go to Step 1, until $j = n$.

5 Experimental Results and Analysis

5.1 The Experiment Data

The experiment data used in this paper is a random segment of data extracted from a substation database. The data records four attributes of network node Temperature, Absolute Pressure, Density and Moisture, totaling 1000 pieces of data. Some of the data are shown as follows (Table 2):

Table 2. Some examples of data used in the experiment

Time	Temperature	Pressure	Density	Moisture
2000/1/2 15:40:54	19.39	0.408	0.409	18
2000/1/8 20:24:53	15.62	0.411	0.418	8.1
2000/1/13 10:46:28	9.42	0.402	0.419	12.8
2000/1/17 9:27:30	1.49	0.376	0.404	10.8
2000/1/20 18:35:13	21.13	0.402	0.4	20.8
2000/1/23 21:42:07	17.82	0.604	0.61	26.2

5.2 The Impact of Individual Data Quality Indicator Improvement on the Overall Data Quality

Table 3 gives the impact of data processing operations on the overall data quality.

Table 3. The impact of data processing operations on the overall data quality

	$D_{\text{integrity}}$	$D_{consistency}$	D_{accuracy}	$D_{timeliness}$	Q
P_0	95.1%	85.4%	88.5%	77.5%	86.6%
$P_{\text{integrity}}$	100.0%	82.1	85.5%	77.5%	86.3%
$P_{consisency}$	83.6%	94.4%	87.8%	77.5%	85.8%
$P_{accuracy}$	86.1%	88.6%	93.7%	77.5%	86.5%
$P_{\text{timeliness}}$	95.1%	85.4%	88.5%	84.8%	88.4%

From the above table, it can be seen that the operation of data quality improvement corresponding to each data quality indicator will inevitably increase the value of the

overall data quality. At the same time, the operation of one data quality indicator will also have a certain impact on other data quality indicators and affect the overall data quality.

5.3 Data Quality Improvement Using the Proposed Method

In this section, the proposed method is used for data quality improvement. To test the performance of the proposed method, certain operations are selected and ordered for comparison, as shown in Table 4, and results are shown in Fig. 2.

Table 4. Some pre-defined operations orders for comparison

Experiment 1	
Order	**Detail**
P_1	$p_{timeliness} \rightarrow p_{consistency}$
P_2	$p_{timeliness} \rightarrow p_{accuracy}$
P_3	$p_{integrity} \rightarrow p_{consistency}$
Experiment 2	
Order	**Detail**
P_1	$p_{timeliness} \rightarrow p_{integrity} \rightarrow p_{accuracy}$
P_2	$p_{consistency} \rightarrow p_{timeliness} \rightarrow p_{accuracy}$
P_3	$p_{integrity} \rightarrow p_{accuracy} \rightarrow p_{timeliness}$
Experiment 3	
Order	**Detail**
P_1	$p_{consistency} \rightarrow p_{accuracy} \rightarrow p_{integrity} \rightarrow p_{timeliness}$
P_2	$p_{consistency} \rightarrow p_{timeliness} \rightarrow p_{integrity} \rightarrow p_{accuracy}$
P_3	$p_{integrity} \rightarrow p_{accuracy} \rightarrow p_{timeliness} \rightarrow p_{consistency}$
Experiment 4	
Order	**Detail**
P_1	$p_{timeliness} \rightarrow p_{accuracy} \rightarrow p_{integrity} \rightarrow p_{consistency} \rightarrow p_{timeliness}$
P_2	$p_{timeliness} \rightarrow p_{accuracy} \rightarrow p_{integrity} \rightarrow p_{consistency} \rightarrow p_{accuracy}$
P_3	$p_{integrity} \rightarrow p_{consistency} \rightarrow p_{integrity} \rightarrow p_{consistency} \rightarrow p_{accuracy}$

From Fig. 2 it can be seen that in the first three groups of comparative experiments, the data quality gain of data processing strategy obtained by greedy algorithm is the largest. In the fourth group of experiments with repeated operations, only the data quality gain of the third path is slightly larger than that of the data processing strategy obtained by greedy algorithm. Therefore, the effectiveness of the proposed algorithm can be verified.

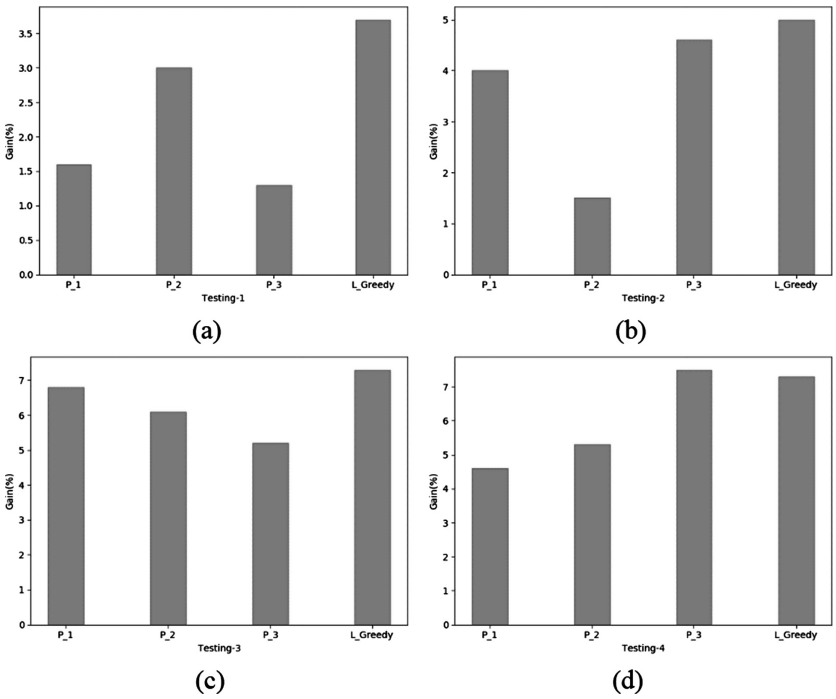

Fig. 2. Comparison of data quality gains from different data processing orders

6 Conclusion

This paper proposes a novel data quality improvement method based on the greedy algorithm. First, this paper establishes four data quality indicators, and gives the calculation formula of data quality. Second, a greedy algorithm is adapted for data quality improvement to find an efficient and reasonable data processing strategy. Finally, simulation experiments prove the correctness of the theorem and the validity of the data processing strategy.

Acknowledgments. This work was supported by the State Grid Corporation Science and Technology Project (Contract No.: SGLNXT00YJJS1800110).

References

1. Li, Cai, Yu, L., Zhu, Y., et al.: Historical evolution and development trend of data quality. Comput. Sci. **45**(4), 1–10 (2018)
2. Saha, B., Srivastava, D.: Data quality: the other face of big data. In: IEEE International Conference on Data Engineering. IEEE (2014)
3. Wang, R.Y., Strong, D.M.: Beyond accuracy: what data quality means to data consumers. J. Manag. Inf. Syst. **12**(4), 5–33 (1996)

4. Sidi, F., Panahy, P.H.S., Affendey, L.S., et al.: Data quality: a survey of data quality dimensions. In: International Conference on Information Retrieval & Knowledge Management (2012)
5. Zaveri, A., Rula, A., Maurino, A., et al.: Quality assessment for linked data: a survey. Semant. Web **7**(1), 63–93 (2015)
6. Wang, Z., Yang, Q.: Research on the quality and standardization of scientific data. Stand. Sci. **03**, 25–30 (2019)
7. Mohan, Li, Li, J., Gao, H.: Solution algorithm for data timeliness determination. J. Comput. Sci. **35**(11), 2348–2360 (2012)
8. Fan, W., Geerts, F.: Relative information completeness. ACM Trans. Database Syst. (TODS) **35**(4), 1–44 (2010)
9. Fan, W., Li, J., Ma, S., et al.: Interaction between record matching and data repairing. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2011, 12–16 June 2011, Athens, Greece. ACM (2011)
10. Fan, W., Ma, S., Tang, N., Yu, W.: Interaction between record matching and data repairing. J. Data Inf. Qual. (JDIQ) **4**(4), 16 (2014)
11. Quercia, D., Hogan, B.: Proceedings of the Ninth International AAAI Conference on Web and Social Media - ICWSM 2015. AAAI Press (2015)
12. Ding, X., Wang, H., Zhang, X., et al.: Research on the relationship among various properties of data quality. J. Softw. **27**(7), 1626–1644 (2016)
13. Cheng, H., Feng, D., Shi, X., et al.: Data quality analysis and cleaning strategy for wireless sensor networks. Eurasip J. Wirel. Commun. Netw. **2018**(1), 61 (2018)
14. Kleindienst, D.: The data quality improvement plan: deciding on choice and order of data quality improvements. Electron. Markets **27**(4), 1–12 (2017)
15. Helfert, M., Foley, O., Ge, M., et al.: Limitations of Weighted Sum Measures for Information Quality (2009)
16. Batini, C., Cappiello, C., Francalanci, C., et al.: Methodologies for data quality assessment and improvement. ACM Comput. Surv. **41**(3), 16 (2009)
17. Zhao, W., Li, C.: A review of the research on quality evaluation methods of associated data. Intell. Theory Practice **39**(02), 134–138+128 (2016)
18. Liu, H.: Analysis of statistical data quality. In: International Joint Conference on Computational Sciences & Optimization. IEEE (2014)
19. Alpar, P., Winkelsträter, S.: Assessment of data quality accounting data with association rules. Expert Syst. Appl. **41**(5), 2259–2268 (2014)
20. Vaziri, R., Mohsenzadeh, M., Habibi, J.: Measuring data quality with weighted metrics. Total Qual. Manag. Bus. Excellence **30**(5–6), 708–720 (2019)