Research Article

Equipment Quality Data Integration and Cleaning Based on Multiterminal Collaboration

Cui-Bin Ji ¹, Gui-Jiang Duan,^{1,2} Jun-Yan Zhou,³ and Wei-Jie Xuan⁴

¹Department of Industry and Manufacturing System Engineering, Beihang University, Beijing, China

²Jingdezhen Research Institute of Beihang University, Beijing, China

³Southwest Technology and Engineering Research Institute, Chongqing 400039, China

⁴Beijing Institute of Control and Electronics Technology, Beijing 100038, China

Correspondence should be addressed to Cui-Bin Ji; geiliwangyi5@163.com

Received 6 May 2021; Revised 14 September 2021; Accepted 9 October 2021; Published 31 October 2021

Academic Editor: Xianming Zhang

Copyright © 2021 Cui-Bin Ji et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the advancement of digital manufacturing technology, data-driven quality management is getting more and more attention, and it is developing rapidly under the impetus of technology and management. Quality data are growing exponentially with the help of increasingly interconnected devices and IoT (Internet of Things technologies). Aiming at the problems of insufficient quality data acquisition and poor data quality of complex equipment, the research on quality data integration and cleaning based on digital total quality management is carried out. The data integration architecture of complex equipment quality based on multiterminal collaboration is constructed. The architecture integrates a variety of integration methods and standards, such as XML, OPC-UA, and QIF protocol. Then, to unify the data view, a cleaning method of complex equipment quality data based on the combination of edit distance and longest common subsequence similarity calculation is proposed, and its effectiveness is verified. It provides the basis for the design of the digital total quality management system of complex equipment.

1. Introduction

Digital total quality management is a management concept driven by technology and management. It is the inevitable choice of enterprise management under the current digital revolution and market competition. The connotation of digital total quality management is as follows.

In terms of management, it is to plan, control, and improve the whole process with the customer as the center; all staff participation as the basis; process management as the main starting point; data statistics and analysis as the basic means; customer loyalty and zero defect as the main goal; the establishment of a digital quality evaluation system as the foundation; and the sharing and utilization of expert knowledge as the basis to implement the planning, control, and improvement of the whole process to satisfy customers. It is a kind of total quality management that benefits the organization's owners, employees, suppliers, partners, or other relevant parties to make the organization achieve long-term success.

In terms of technology, it is based on modern information technology, with modern quality management technology as the core, combined with cloud technology, Internet of things technology, intelligent manufacturing technology, equipment health management and prediagnosis technology, big data technology, and other new technologies, which are comprehensively applied to all stages of enterprise product design, manufacturing, test, and service life cycle quality management. According to the automatic real-time collection, transmission, analysis, and feedback control, as well as the sharing of quality information resources and the coordination of quality management, a set of total quality management characterized by digitization, integration, networking, and collaboration, and the combination of software and hardware is established.

Quality management depends on data. "Speaking with data, making decisions with data, relying on data management and innovating with data" is the focus of follow-up quality management. However, the traditional quality

management of complex equipment still has the following problems in quality data integration and cleaning:

- (1) At present, most quality management systems focus on the analysis and processing of management quality information, while the quality data management in the production process under discrete manufacturing mode is slightly lacking
- (2) The quality data collection in the production process is not standardized, there are many types of NC equipment involved in the workshop, and the data protocols of the equipment are complex and diverse, resulting in the diversity of quality data collected and poor interconnection and interoperability
- (3) There is no effective data cleaning method for multisource, heterogeneous, and massive quality data

To solve the above problems, scholars at home and abroad have made relevant research in data integration, data interconnection, and interoperability and data fusion processing. Aiming at the problem of quality data integration in production process, Lei et al. [1] proposed a manufacturing process quality management method based on manufacturing bill of materials (MBOM) and studied the organizational structure and management of QBOM (Quality BOM). According to the current situation of disordered and inefficient quality information management in the manufacturing process of domestic automobile manufacturing enterprises, Kuang Zhuping et al. [2] developed a quality management system for automobile manufacturing process and elaborated the business requirements and business processes of quality management in manufacturing process. In view of the compilation and management of inspection planning, Yan and Tang [3] analyzed the working principle and functional model of CAIP and proposed a CMM integration scheme based on DMIS standard.

Aiming at the interconnection and interoperability of workshop quality data, MTconnect [4] introduces a communication connection protocol for CNC machine tools, which aims to enhance the interoperability among CNC systems, equipment, and application software. Chen et al. [5], Qin [6], and Zhang [7] established a unified semantic information model for typical manufacturing resources based on OPC-UA standard protocol and realized data collection and unified management of typical equipment in workshop. Ning [8] proposed an embedded OPC-UA wireless sensor network data acquisition and monitoring architecture and realized wireless sensor network equipment monitoring. In terms of quality data collection and management, Zhao et al. [9] introduced the basic principle and application method of QIF and pointed out that QIF overcame the shortcomings of previous manufacturing quality system related standards. Bernstein et al. [10] pointed out that QIF provides a standardized method for quality information collection and exchange in the field of measurement and that QIF adopts extensible markup language (XML) architecture mode to facilitate integration with other

industrial information standards. The above research is mainly based on the standard protocol to collect and manage the data, but there are few modeling of the perception equipment for collecting the workshop quality data, and the research on the quality data standardization based on QIF is also less. Therefore, it is necessary to collect the perception equipment and quality data of the manufacturing workshop based on the standard protocol.

In the research of complex equipment quality data integration architecture, reference [11] introduced data warehouse technology into quality information management and established quality information multidimensional data warehouse model. In order to achieve decision support, the massive data processing modeling is studied. In reference [12], a quality traceability management platform was constructed by using middleware technology based on layered technology. Different functions of all kinds of equipment are shielded, and redundant data are filtered based on corresponding protocol specifications to realize real-time quality information traceability. References [13–15] designed the corresponding quality data acquisition scheme based on the Internet of things technology in the manufacturing workshop. In terms of data cleaning, reference [16] proposed a data cleaning model for product quality standards. Reference [17] proposed a perceptual data cleaning method based on spatiotemporal correlation. Reference [18] proposed a data cleaning and fusion method based on time series for distribution network data. Reference [19] constructed a data cleaning and repairing model based on improved GMM algorithm.

Aiming at the shortcomings of the existing research, this paper carries out the research on the quality data integration and cleaning of complex equipment. Implement the concept of quality data as the center, and put forward a complete solution of quality data acquisition based on multiterminal cooperation, quality data integration based on ODPS, and quality data cleaning based on algorithm. It realizes the centralized collection and control of quality data in the whole life cycle of complex equipment, efficiently cleans the noisy quality data and unifies the quality data view, provides a solution for the interworking of quality data of complex equipment, and lays a foundation for the integration of digital total quality management system.

2. Multiterminal Collaborative Overall Framework Design

The overall plan is constructed with the idea of synergy between the cloud, the device side, the management side, and the system integration side.

The overall framework of the design is shown in Figure 1. The advantages of the cloud are mainly reflected in the following two points: first, abundant computing resources can quickly perform a large number of complex calculations. Second, sufficient storage resources can store massive samples for training. But at the same time, due to the poor time sensitivity of the cloud and poor task customization capabilities, it often fails to meet real-time requirements. The device side, management side, and system side have two

FIGURE 1: Overall frame design.

advantages because they are tightly coupled with application objects: first, good real-time responsiveness and second, the service object being single, and personalized service can be customized. However, its computing resources often become a bottleneck restricting large-scale data-driven computing. Based on the above analysis of the advantages and disadvantages of each terminal, this article gives full play to the respective advantages of each terminal.

In Figure 1, ODPS mainly carries out the storage of quality data and the training of cleaning algorithms. Among them, the quality data acquisition end consists of three parts: ① distributed quality dataset at the integrated system end, ② quality dataset collected at the management end, and ③ real-time quality dataset collected at the equipment end. Then, based on the rich quality data sample resources and computing resources in the cloud, the quality data cleaning model is continuously trained and updated to obtain

universal training results, which can serve different analysis scenarios as intermediate results.

In the integrated system, ① the integration method based on Web Service and RESTful is developed to collect the quality data in relevant systems and upload them to the cloud for storage.

In the management side, ② the standardized process collection and result presentation of documents generated in the process of production and operation, such as technical status documents and product quality certificates, are carried out by formulating collection templates and work specifications and transmitted to the cloud storage.

In the device end, ③ the sensing data under personalized working conditions, such as real-time status data of environment, are standardized through the standard OPC-UA server through the sensing device and are transmitted to the cloud for storage; the measurement equipment standardizes

the measurement data related to parts through the QIF server and transmits it to the cloud for storage.

It can be seen from Figure 1 that, in the process of quality data integration, the data transmitted from the edge multiend to the cloud mainly include two things: first, the device will regularly upload the collected quality data to the cloud storage through standardized processing to enrich the cloud training set and verification set. Second, the management end and the integration end will periodically upload the local quality data to the cloud storage and complete the optimization of the data cleaning algorithm parameters through training regularly, and the optimized algorithm will clean the quality data.

3. Quality Data Collection Based on Multiterminal Collaboration

3.1. Perceived Quality Data Collection Based on Standard Protocol Model. The interoperability of heterogeneous equipment in workshop is the basis of intelligent manufacturing, and the interoperability of equipment must require its interconnection and data sharing. Therefore, the multisource heterogeneous data between devices must be able to realize interoperability, and the unification of data protocol is the basis of data interconnection. The internationally popular industrial equipment communication protocols include MTconnect and OPC unified architecture (OPC-UA), and the quality data standard protocol includes quality information framework (QIF) standard. Therefore, in this paper, OPC-UA standard protocol is used for semantic modeling of workshop perception equipment, and QIF standard protocol is used for semantic modeling of workshop typical measurement equipment, so as to study and implement the data acquisition method of workshop typical equipment based on OPC-UA and QIF.

3.1.1. OPC-UA Semantic Model of Perception Device. The key to realize the interconnection between sensing devices is to abstract the physical devices and form the information model of the devices, so that a type of equipment presents a standard interface to the outside world and establishes the communication specification on this basis. OPC-UA is becoming a potential solution for modeling and formulating communication specifications of public equipment due to its unique advantages.

The OPC-UA semantic model is created according to the quality information related standards and specifications. Taking a typical sensing device as an example, the main components of its object type include two parts: sensor object and RFID identifier object. These two objects are created according to their corresponding object types, and each object inherits all the data structures of its object type.

In order to better express the environmental status of the workpiece and the internal relationship of traceable product information and provide effective data support for the subsequent data fusion, this paper only lists the sensor and RFID identifier in the sensing equipment and their related data items and uses OPC-UA modeling rules to model.

Therefore, the established OPC-UA semantic model of the sensing device is shown in Figure 2, through which the effective monitoring of the working status of the sensing device can be realized.

(1) Sensor Type Instantiation. At present, there are three instance objects of sensor types in the system, namely temperature, humidity, and dust sensor. The sampling interval of the three sensors is 100 ms by default. The initial default value of the sensor's reading variable is set to 0. Once the real-time data source is bound to the corresponding node of the instantiated information model, the sensor's reading will be automatically updated to the actual value. The same is true for the subsequent object type instantiation. All node variables bound to the underlying data are set to 0, and string type variables are set to empty. The instantiation of the sensor type is shown in Figure 3.

(2) RFID Reader Type Instantiation. Figure 4 shows the instantiation of the RFID recognizer type. The default time between RFID readers to read the card is 500 ms. Once the data information is read, the EPC variable will display the latest code read. The EPC code length is 96 bits, which is represented by a hexadecimal number.

In OPC-UA information modeling, in order to visualize the information model, UML-like graphics are used to represent eight node categories and some commonly used reference types, so that the information model can be displayed intuitively. However, in practical applications, the computer cannot understand the meaning of the graphic representation, so it must adopt a form that the machine can parse. A very suitable solution is to express the information model through an XML file.

XML is an extensible markup language developed by the International Standards Organization W3C for marking data and defining data types. It uses extensible and self-describing features to structure data and documents so that computers can understand them. XML supports data sharing between different vendors or applications. With the rapid development of network applications such as e-commerce and web services, XML has become the most widely used identification technology and has been supported by many programs and libraries.

After the design of the information model is completed, an XML model conforming to the OPC-UA regulations can be automatically generated, as shown in Figure 5. In this way, an XML document that contains sensor type information and can be parsed by the OPC-UA server is generated.

3.1.2. QIF Semantic Model of Measuring Equipment. The QIF standard covers a variety of use cases, including dimensional metrology inspection, first article inspection, reverse engineering, and discrete quality measurement. This section mainly uses QIF standards to model measurement equipment (including three coordinate measuring machines, articulated arm measuring machines (standard semantic conversion can be achieved through external

FIGURE 2: Perceptual device information model.

FIGURE 3: Sensor type instantiation.

references), and measuring instruments such as vernier calipers and micrometers) and their measurement data information, as shown in Figure 6.

Inspection equipment is the most important resource in inspection activities in manufacturing companies. The inspection information model includes (1) the basic types of

FIGURE 4: RFID identifier type instantiation.

FIGURE 5: Sensor type information.

equipment, such as CMM and digital calipers; (2) manufacturer information; (3) occupied volume; (4) calibration information, used to describe the specific calibration requirements information of the inspection equipment; (5) use environment, used to describe the range of environmental conditions for the normal operation of the inspection equipment, such as temperature, air pressure, humidity, and vibration; and (6) purchase time. The information contained in software resources is divided into three categories: (1) descriptive information (DeIn), resource unique identification, resource name, resource type, resource version, and download path; (2) resource preparer information (owner), including the preparer, preparer ID, preparer's department, and preparer's contact information; and (3) function index (AbIn), including application environment, instructions for use, and supplementary information. Human resources information includes two major categories: basic information and organization information. The human resources information is filled in when the personnel register, and the human resources department maintains it. Basic Information (Basic Inf) fills in the personal information of an

employee, such as personal name, ID, contact information, identity or job information, and skill information. Organization Information (Organization Inf) contains the department information to which an employee belongs, such as department name, department ID, department size, and necessary supplementary information.

3.2. Specification-Based OCR Acquisition Model. The recognition accuracy rate of OCR has a limit characteristic: it can only be infinitely close to a certain value but never reach it. The main reason is that there are too many factors that affect the recognition rate: differences in writing habits, scanner scanning quality, document printing quality, recognition methods, learning and testing samples, memory vocabulary size, etc.

Compared with the United Kingdom and the United States, the development of OCR technology in China has not been very smooth. The number of Chinese characters and the number of English letters are not of the same magnitude. There are many strokes in Chinese characters, the writing

FIGURE 6: QIF information model of measurement data.

style is very random, and there are a large number of similar characters, such as “人” and “入”. Therefore, determining the writing specifications and process specifications of industry documents can greatly reduce the difficulty of OCR identification.

3.2.1. Template Creation Process. According to the relevant documents of national standards and corporate standards, organize and refine the structured description of the item information, format information, content name information item list, and the corresponding relationship of the above information list to form a template, and the template process is shown in Figure 7.

3.2.2. Work Process Specification. The basic workflow method of document list recognition based on OCR technology is proposed, including image import, image preprocessing, comparison recognition, modification and correction, results collation and output, and results acceptance, as shown in Figure 8.

Before the implementation of file OCR, it is necessary to evaluate whether the quality of digital copies of paper files meets the basic requirements of OCR. The assessment content should generally include image resolution, skewness, sharpness, distortion, brightness, contrast, and grayscale. Digital copies of paper archives whose quality cannot meet the basic requirements of OCR work should be redigitized and imported in accordance with the requirements of DA/T31. The image preprocessing of the document list mainly includes binarization, image noise reduction, tilt correction, and image detection.

3.3. Integration Model Based on Open Data Processing Service. Open data processing service (ODPs) is a platform as a service (PAAS) platform produced by Alibaba. It can provide big data acquisition, cleaning, management, and analysis capabilities; support the standardization and rapid customization of business applications; help reduce the possibility of data I/O throughput, unnecessary data redundancy, and data error; realize the reuse of calculation results; and improve the efficiency of data utilization. In order to avoid data chimney and inconsistency and realize standardized, shared, and reusable data integration, combined with the characteristics of equipment quality data and business practice, the data integration architecture of on-site equipment quality data based on ODPS is designed, as shown in Figure 9.

The digital total quality management system integration platform implements the concept of centering on quality data, adopts the authorization and credit mechanism and means, and adopts the electronic signature authentication strategy to ensure the high credibility of the business processing operation data. The equipment quality data collection architecture is shown in Figure 2. The data collected by sensor devices, RFID, and other sensing devices are collected. MiNiFi agents use point-to-point protocols (such as HTTP/HTTPS) to transfer the collected data to the central

Kafka cluster. The state data that needs real-time processing is stored in the real-time distributed database such as H-Base; the non-real-time data such as OCR technology and external system integration process are stored in distributed file systems such as hive distributed file system (HDFS).

In the end, all data will be stored in a distributed file system for extraction. By subscribing to the data of the corresponding topic (topic), users can use a processing framework such as Spark Streaming [20] for big data processing to meet the requirements of equipment support business applications. In addition, the cluster server high-availability design [21] can prevent the phenomenon of data transmission being affected by a server downtime.

Because the data items that need to be monitored in the industrial field are very large, the field collection system will collect a large amount of production data at a high frequency, which requires a large amount of network bandwidth during transmission. The data collection cluster is deployed on the cloud server to achieve efficient data transmission. In this system, the data collection cluster in the cloud server uses Rabbit MQ cluster and Kafka cluster to collect non-real-time data and real-time data, respectively, and store all data in the database cluster of the industrial cloud, that is, to transmit the live real-time data to the industrial cloud computing center. As for Kafka real-time data collection cluster, when the network is interrupted, the historical data cached in the local industrial database is transmitted to the Rabbit MQ non-real-time data collection cluster of the industrial cloud and stored in the cloud database, thereby avoiding the problem of incomplete data caused by the network interruption. At the same time, it installs and supports distributed system data synchronization services, and high-availability services are formed between cluster servers. When one server fails, another server can be used to provide services to prevent data transmission from being affected by server failure.

4. Quality Data Resource Cleaning Algorithm

All quality business and management data are unified in the cloud database through data collection services. Due to the different standards and specifications of data fields and the strong randomness of manually reporting data, there are many different forms of expression for data that express the same meaning. During data integration, in order to unify the view, these data in different manifestations need to be correctly cleaned into the target data, and unified and standardized conversion is performed. Therefore, the effective cleaning of the quality data collected in the data center has become one of the key problems of quality data integration.

At present, the repeatability calculation method is used to clean the quality data. The classification accuracy of this cleaning is not high. The specific performance is that some data are not classified into any category in the classification target because the repeatability cannot reach the threshold value, and some data are classified into the wrong category. Therefore, according to the actual characteristics of quality data, this paper proposes a quality data resource cleaning

FIGURE 7: Collection template creation process.

FIGURE 8: Work standard process.

FIGURE 9: Integration framework based on ODPS.

method based on similarity calculation of editing distance and longest common subsequence and residual similarity calculation, in order to improve the cleaning accuracy.

4.1. Quality Data Resource Cleaning Model Construction.

The general data cleaning model mainly includes incomplete data processing, redundant data processing, and error data processing. Quality data have the characteristics of being multisource and heterogeneous, including structured, unstructured, semistructured, and time series data. According to the characteristics of quality data resources, the quality data without the support of complete scientific theory system cannot correct the error data, cannot eliminate the wrong data, and cannot find the data, which is defined as error or error. They are all quality data resources obtained in the whole life cycle. Even in some cases, redundant data processing cannot be carried out, because multiple redundant repetitions correspond to multiple occurrences, which helps to strengthen the weight and facilitate subsequent analysis. Aiming at the polymorphic and complex characteristics of quality data resources, this paper constructs a quality data cleaning model that integrates incomplete data processing algorithms and redundant data processing algorithms and corrects and eliminates quality data resources in the manufacturing process and management process, so as to output high-quality data.

4.1.1. Data Loading and Preprocessing. It is necessary to read the database resources into the program. For the table structure relationship of the current relational database, the two-dimensional linked list is used to extract and load the data, so as to provide fast loading, modification, reading, and deletion of rows and columns.

Except for time and date data, other data such as numeric and text data can be directly used for calculation, so special preprocessing is performed on time and date data. The 10 bit time stamp is used to process the time, and the time stamp can be multiplied and optimized here. For example, if a field records the service life of the lathe, there is no need for a 10 bit time stamp, because the 10 bit time stamp is accurate to second level data and can be rounded to the first 7 bits of recorded data, that is, the time interval close to 20 minutes. If the requirements are more extensive, you can even record the time interval trend of the top 6, that is, about 3 hours. In particular, due to the subsequent calculation of residual similarity for the digital data, the residual results are basically unchanged for the calculation method of narrowing and expanding at the same time (rounding may lead to slight changes). In this case, the purpose of the calculation is to change the result of the calculation instead of changing the result of the calculation.

4.1.2. Text Similarity Calculation

(1) *Similarity Calculation Based on Edit Distance.* Levenshtein Distance (LD) is a string similarity calculation method. The minimum number of times a string becomes another string is for insertion, change, and deletion, which is usually calculated by dynamic programming method.

Given two strings string1 and string2, $DP[i, j]$ represents the minimum edit distance between the first i characters of string1 and the first j characters of string2, $i \in [0, \text{length}(\text{string1})]$, $j \in [0, \text{length}(\text{string2})]$. In particular, it is initialized when $ij = 0$, such as $i = 0, j = 3$ means that string1 takes the first 0 characters, that is, null to string2 takes the first 3 characters, and the minimum number of operations is equal to 3.

Initialization:

$$DP[i, j] = i + j, \quad ij = 0. \quad (1)$$

Recursive formula:

$$DP[i, j] = \begin{cases} \min(DP[i-1, j-1], \min(DP[i-1, j], DP[i, j-1])) + 1, & ij > 0, \text{string1}(i) \neq \text{string2}(j), \\ DP[i-1, j-1], & ij > 0, \text{string1}(i) = \text{string2}(j). \end{cases} \quad (2)$$

In the recursive formula, $\text{string1}(i)$ represents the i -th character of string1 . If $i=0$, it is null, and $\text{string2}(j)$ is the same.

Use edit distance to calculate text similarity in quality data resources. Because of its data characteristics, it is accurate and precise. For example, the data of recording location “x factory C8 lathe” and “x factory D9 lathe” are the edit distance when calculating similarity. Compared with “x factory C8 lathe” and “y factory D9 lathe”, the editing distance is shorter, which proves that the editing distance is reasonable to distinguish the location and include the semantics of the location, and the similarity of “x factory D9 lathe” and “y factory D9 lathe” compared with “x factory C8 lathe” and “y factory D9 lathe” has a shorter editing distance, which proves that the editing distance is reasonable to divide the model and incorporate the semantics of the model. Edit distance makes reasonable use of semantic information and fits with quality data resources.

For example (Table 1), calculate the similarity of the two pieces of data “x factory C8 lathe” and “x factory D9 lathe”, and finally take $DP[\text{length}(\text{string1}), \text{length}(\text{string2})]$ as the final result; that is, the minimum edit distance is 2.

$$DP[i, j] = \begin{cases} \max(DP[i-1, j-1], \max(DP[i-1, j], DP[i, j-1])) + 1, & ij > 0, \text{string1}(i) \neq \text{string2}(j), \\ DP[i-1, j-1], & ij > 0, \text{string1}(i) = \text{string2}(j). \end{cases} \quad (4)$$

In the recursive formula, $\text{string1}(i)$ represents the i -th character of string1 . If $i=0$, it is null, and $\text{string2}(j)$ is the same.

In the quality data resource cleaning, the text similarity is calculated based on the longest common subsequence, because the longer the two text repeat subsequences, the closer the data is, the more similar it is. For example, record the data of lathe parts “x lathe y parts”, which is stored as “X Window y parts” due to workers’ misoperation. When calculating the longest common subsequence, the two data are 5, while the length of both data is 6, so the similarity of the two data is very high. The longest common subsequence is more widely used in Chinese texts, because the basic unit of Chinese is Chinese characters, and Chinese characters themselves contain certain semantic information. Therefore, using the longest common subsequence to calculate text similarity can obtain reasonable similarity calculation at the semantic level.

For example (Table 2), calculate the similarity of the two pieces of data “x lathe y part” and “x car window y part”, and finally take $DP[\text{length}(\text{string1}), \text{length}(\text{string2})]$ as the final

(2) *Calculation Based on the Longest Common Subsequence Similarity*. The longest common subsequence (LCS) is a string similarity calculation method. The common subsequence of the subsequence is calculated without considering the relative position of the characters. The more repeated subsequences indicate more similarity. Usually use dynamic programming method for calculation.

Given two strings string1 , string2 , $DP[i, j]$ represents the maximum substring length from the first i characters of string1 to the first j characters of string2 , $i \in [0, \text{length}(\text{string1})]$, $j \in [0, \text{length}(\text{string2})]$. In particular, when $ij=0$, it is initialized. For example, $I=0, j=3$ means that string1 takes the first 0 characters, that is, null to string2 , and the maximum repeated substring length of the first three characters is equal to 0.

Initialization:

$$DP[i, j] = i + j, \quad ij = 0. \quad (3)$$

Recursive formula:

result; that is, the longest common subsequence similarity is 5.

(3) *Joint Similarity Calculation Based on Edit Distance and Longest Common Subsequence*. The text similarity calculated based on the edit distance and the longest common subsequence has a certain semantic analysis meaning, but the following problems need to be solved. Hereinafter, the similarity calculation based on the edit distance is called LD, and the corresponding calculated value is $DP(LD)$, based on the longest LCS, and the corresponding calculated value is $DP(LCS)$.

- (1) The smaller the LD result, the higher the similarity. When $DP(LD)=0$, the two texts are completely similar; the larger the LCS result, the higher the similarity. When $DP(LCS)=0$, the two texts are literally completely dissimilar and need to get a reconciled calculation.
- (2) The results of $DP(LD)$ and $DP(LCS)$ are both numbers, not a reasonable ratio, because a ratio is

TABLE 1: Schematic diagram of recursive process of LD.

	0	X 1	Workshop 2	C 3	8 4	Turning 5	Machine 6
X	1	0	1	2	3	4	5
Workshop	2	1	0	1	2	3	4
D	3	2	1	1	2	3	3
9	4	3	2	2	2	2	3
Turning	5	4	3	3	3	2	3
Machine	6	5	4	4	3	3	2

TABLE 2: Schematic diagram of recursive process of LCS.

	0	x 0	Turning 0	Machine 0	y 0	Spare 0	Part 0
X	0	1	1	1	1	1	1
Turning	0	1	2	2	2	2	2
Window	0	1	2	2	2	2	2
Y	0	1	2	2	3	3	3
Spare	0	1	2	2	3	4	4
Part	0	1	2	2	3	4	5

obtained through reconciliation; otherwise, the result of the long string and the result of the short string are too different, and this ratio is the best reconciliation to the interval [0,1], representing the percentage. In comprehensive analysis, the first typical algorithm is to find the reciprocal of DP (LD) or DP (LCS), a unified measurement, and both are in the same direction or reverse scheduling with the similarity, finding the harmonic average (the sum of the reciprocal of 1 ratio):

$$\text{similarity} = \frac{1}{(1/\text{DP}(\text{LCS})) + \text{DP}(\text{LD})} \quad (5)$$

$$= \frac{\text{DP}(\text{LCS})}{1 + \text{DP}(\text{LD}) \bullet \text{DP}(\text{LCS})}$$

Calculation results of this method are as follows:

- (i) When completely similar, similarity = DP (LCS), similarity is still closely related to length, and when completely dissimilar, similarity = 0. It is not mapped to the interval [0, 1] and is closely related to the length of the text itself.
- (ii) To meet the universal similarity change, the larger the DP (LCS), the greater the total similarity, and the smaller the DP (LD), the smaller the total similarity.

A more reasonable similarity calculation method is used:

$$\text{similarity} = \frac{\text{DP}(\text{LCS})}{\text{DP}(\text{LD}) + \text{DP}(\text{LCS})} \quad (6)$$

The calculation result at this time is given as follows:

- (i) When completely similar, similarity = 1, and when completely dissimilar, similarity = 0. Mapped to the interval [0, 1], it has no relation to the length of the text itself.

- (ii) To meet the universal similarity change, the larger the DP (LCS), the greater the total similarity, and the smaller the DP (LD), the smaller the total similarity.

In particular, if a certain text is null, record similarity = 0.

4.1.3. Numerical Similarity Calculation Based on Residuals.

Numerical similarity calculation formula is given as follows:

$$\text{similarity} = 1 - \alpha \frac{\text{abs}(\text{num1} - \text{num2})}{\max(\text{abs}(\text{num1}), \text{abs}(\text{num2}))} \quad (7)$$

In the above formula, abs() represents the absolute value calculation, max() takes the maximum value, num1 and num2 refer to two numbers used for comparison, and α represents the weight coefficient, and the default is 1.

Using the idea of residual error, the residual error rate is calculated based on a larger value, and 1 minus the residual error rate is used as the calculated similarity value. The more closer to 1, the more similar.

For quality data resources, the significance of digital residual calculation results is quite different. For example, the tolerance of a machine tool is 100 ± 0.001 cm, the actual maximum tolerance of machine tool A is 100.0002, and that of machine tool B is 100.0003. If the numerical similarity is calculated, the similarity is approximately equal to 1, which is completely similar. However, on the tolerance level, there is a big difference between 0.0002 and 0.0003. For this type of quality data resources, the final results are calculated after reasonable processing.

For different quality data resources, the digital type needs to increase the weight α calculation, which is an empirical number, which describes the proportion of different column data. For example, the actual working size of a machine tool may be between [2000 mm, 2100 mm], and the calculated residual rate is within 5%. If the weight coefficient is 1, the similarity is always greater than or equal to 0.95. In some cases, it is unreasonable, so the rationality of residual calculation is adjusted by weight coefficient. For the time stamp data to calculate the residual rate, the weight coefficient can also be used to change the rationality of data similarity calculation.

In particular, if num1 or num2 = 0, skip this field.

Data cleaning involves the core calculation similarity formula as follows:

$$\text{similarity_item} = \frac{\sum \text{similarity}(\text{item1}_i, \text{item2}_i)}{n} \quad (8)$$

where item1\item2 refers to two different data, item1_i refers to the i -th data field in the data of Item1, similarity refers to the similarity calculation result of two different data, and n is the total number of data fields, $i \in [1, n]$.

4.1.4. Redundant Data Deduplication. A typical problem of quality data resources is the existence of redundant data, which needs reasonable deduplication. For example, there may be two very close fault records in fault record data. This type of quality data resources cannot be redundant and duplicated. For the machine tool account, redundant

deduplication is needed, because there may be manual record redundancy, which needs to be cleaned. Therefore, before deduplication cleaning, first determine whether the data needs to be deduplicated and whether it can be deduplicated. In many cases, the quality data resources do not deduplicate, which is equivalent to some duplicate data, which can effectively increase the weight of duplicate data.

Algorithm idea: if $\text{similarity_Item} > \beta$, Item2 will be deleted. β is the threshold parameter. If the similarity is higher than β , it is considered that the two data are too similar or even identical, so it is necessary to delete them again.

4.1.5. Completion of Incomplete Data. A typical problem of quality data resources is data incompleteness. Due to the influence of business and other changes, there are large changes in data format and field, resulting in a large number of incomplete data stored in the database. Generally speaking, incomplete data need to be supplemented when cleaning data, which is of great significance for subsequent quality data resource utilization.

Algorithm idea: traverse all the complete data, and record the maximum similarity_item of the incomplete data and the complete data and its corresponding complete data item. In particular, the most complete data will be filled with incomplete data.

4.2. Quality Data Resource Cleaning Result Evaluation.

Experimental environment: the operating system is windows 10 64 bit operating system, the processor is Intel Core i7-4790s, cpu3.20 ghz, memory 16 GB ram, development tool, idea 2019.1. In order to verify the efficiency of the proposed cleaning algorithm, based on the quality data cleaning process proposed in this paper, the data cleaning algorithm is compared with the data cleaning technology based on *N*-Gram algorithm and SACS-GS algorithm through three factors: recall, precision, and running speed.

Recall is an important index to judge the quality of data cleaning technology. The higher the recall, the more similar and repeated records checked, which can represent the performance of data cleaning technology to a certain extent. Figure 10 shows the recall curve. As can be seen from Figure 10, in terms of recall, the algorithm proposed in this paper has greater advantages than *N*-Gram algorithm and SACS-GS algorithm, and with the continuous increase of the amount of data, the recall of the three algorithms is increasing.

The precision rate is an important index to judge the quality of data cleaning technology. It refers to the proportion of similar repeated records detected, which can truly reflect the quality of an algorithm. Figure 11 is the precision rate curve. As can be seen from Figure 11, the algorithm proposed in this paper has greater advantages than *N*-Gram and SACS-GS algorithms. Although this advantage slows down with the continuous increase of data volume, the advantages of this algorithm can be seen in combination with the advantages of recall.

FIGURE 10: Recall ratio.

FIGURE 11: Precision ratio.

With the continuous development of information and the continuous increase of data records, the running time is also used as the standard to judge the quality of an algorithm. This paper compares the running time of the two algorithms, and the results are shown in Figure 12. As can be seen from Figure 12, when the order of magnitude is low, the advantages of this algorithm over the other two algorithms are not obvious. With the continuous increase of the order of magnitude, the running time of *N*-Gram and SACS-GS algorithms increases rapidly, exceeding the running time of the data cleaning algorithm proposed in this algorithm.

In view of the particularity of quality data resources, in order to verify the effectiveness of the proposed algorithm, we used quality data from an aerospace company in the production process of a certain component: a total of 2000 pieces were used as a dataset for result evaluation and analysis. This is a classified dataset (each quality datum has a corresponding quality evaluation result label). If the predicted result is consistent with the actual result label, it is considered that the classification is accurate; otherwise, it is considered that the classification is wrong.

Each time, 300 pieces of data were taken for training verification and 10 times of results were obtained. The results were cross validated to improve the significance of universal proof. The data after duplication cleaning, data after completion of cleaning, and data after complete cleaning (i.e., the data of two parts of cleaning with deduplication and

FIGURE 12: Runtime comparison.

FIGURE 13: Comparison of C4.5 decision tree classification results.

FIGURE 14: Cart decision tree classification results.

replenishment) were compared with the original data, and four longitudinal dimensions were compared. Since the cleaned data cannot be directly evaluated, the classification algorithm needs to be used to output the results. Therefore,

C4.5 (Figure 13), CART decision tree (Figure 14), and random forest algorithm (Figure 15) are used to compare the vertical dimensions of the three algorithms. The evaluation index is based on the classic machine learning evaluation

FIGURE 15: Random forest classification results.

TABLE 3: Classification result comparison.

	Precision	Recall	Macro average
C4.5 algorithm	0.8	0.812	0.806
Cart algorithm	0.828	0.842	0.834
Random forest algorithm	0.914	0.917	0.915

index. The specific evaluation indexes are P (Precision) and R (Recall). It can be seen from the figure that the data accuracy and recall rate after complete cleaning have been significantly improved.

Compared with *N*-Gram and SACS-GS, it shows the efficiency of this algorithm. After the subsequent classification model verification, the quality data resource cleaning based on the similarity calculation of editing distance and longest common subsequence and residual similarity calculation performs well on the quality data resource dataset and has high cleaning efficiency. The cleaned quality data are compared in horizontal dimensions by using a variety of classification algorithms. The random forest algorithm has obvious advantages over the other two classification methods, as shown in Table 3. It lays a good foundation for the subsequent analysis and prediction reliability and shows that the cleaning of quality data resources is effective and feasible.

5. Conclusion

This paper designs a quality data integration framework based on multiterminal collaboration, which can effectively realize the centralized collection of quality data in the workshop equipment side, management process and external system, and process according to the business requirements. On this basis, aiming at the problem of complex equipment quality data cleaning, combined with the complex and professional characteristics of complex equipment quality data, a cleaning algorithm model based on edit distance and the longest common subsequence similarity calculation is constructed, which effectively improves the accuracy of complex equipment quality data cleaning. In the next step, the application of complex equipment quality data mining will be studied.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] L. Lei, Z. M. Zhang, and T. C. Tian, "Quality management of aircraft manufacturing process based on MBOM," *Aeronautical Manufacturing Technology*, vol. 07, pp. 74–76, 2008.
- [2] Z. P. Kuang, J. X. You, and X. M. Du, "Research and development of process quality management information system for automaker," *Industrial Engineering and Management*, vol. 02, pp. 43–47, 2005.
- [3] H. J. Yan and X. Q. Tang, "Computer aided product inspection planning and design," *China Quality*, vol. 07, pp. 24–26, 2002.
- [4] P. Lei, L. Y. Zheng, L. H. Wang, Y. H. Wang, C. Li, and X. C. Li, "MTConnect compliant monitoring for finishing assembly interfaces of large-scale components: a vertical tail section application," *Journal of Manufacturing Systems*, vol. 45, no. 121–34, 2017.
- [5] Y. Z. Chen, *Research of Configurable Acquisition System for Machine Running Status Data Based on*, Chongqing University, Chongqing, 2018.
- [6] R. K. Qin, *Research on Data Unified Interface of NC System Based on OPC UA Specification*, Huazhong University of Science & Technology, Wuhan, 2017.
- [7] C. E. Zhang, *Research and Implementation of OPC UA Information Model towards Internet of Manufacturing Things*, Huazhong University of Science & Technology, Wuhan, 2016.
- [8] Y. Ning, *Design and Implementation of Industrial Wireless Sensor Network Supporting OPC UA*, Beijing Jiaotong University, Beijing, 2018.
- [9] Y. F. Zhao, J. A. Horst, T. R. Kramer, W. G. Rippey, and R. Brown, "Quality information framework-integrating metrology processes," in *Proceedings of the 14th IFAC Symposium on Information Control Problems in Manufacturing (INCOM)*, Bucharest, RO, May 2012.
- [10] W. Z. Bernstein, T. D. Hedberg, M. Helu, and A. B. Feeney, "Contextualising manufacturing data for lifecycle decision-

- making,” *International Journal of Product Lifecycle Management*, vol. 10, no. 4, p. 326, 2017.
- [11] X. B. Gao, *Research on Key Technologies of Quality Information System Based on Data Warehouse*, Northwestern Polytechnical University, Xi'an, 2005.
 - [12] Z. Li, *The Design and Development of an Platform for Product Quality Traceability of Enterprises*, Hunan University, Hunan, 2018.
 - [13] J. Lv, *Research on Data Acquisition and Processing System of Stamping Workshop Based on Internet of Things*, Jiangsu University, Jiangsu, 2019.
 - [14] X. Liu, W. L. Sun, and Q. W. Cui, “The real-time information extraction and control of workshop manufacturing system based on IOT,” *Modular Machine Tool & Automatic Manufacturing Technique*, vol. 03, pp. 154–157, 2016.
 - [15] G. Song, Y. P. Yuan, and X. J. Li, “Study on the production of plant data acquisition system based on Internet of things,” *Equipment Manufacturing Technology*, vol. 12, pp. 14–16, 2014.
 - [16] Z. J. Wang and L. W. Yue, “Research on data cleaning model of product quality standard and its application,” *Standard Science*, vol. 04, pp. 88–95, 2020.
 - [17] T. C. Sun, *Based on Spatial-Temporal Correlation Sensory Data Cleaning Research*, Beijing University of Civil Engineering and Architecture, Beijing, 2020.
 - [18] Z. Youchan, W. X. Liang, and Y. Z. Wang, “Research on data cleaning and fusion in distribution power grid based on time series technology,” *Power System Technology*, pp. 1–8, 2021.
 - [19] L. Jing, *Research and Application of EM Optimization Algorithm Based on GMM*, Harbin Engineering University, Harbin, 2018.
 - [20] S. Chintapalli, D. Dagit, B. Evans et al., “Benchmarking streaming computation engines: storm, flink and spark streaming,” in *Proceedings of the IEEE International Parallel & Distributed Processing Symposium Workshops*, Chicago, IL, USA, May 2016.
 - [21] Y. M. Hou, Q. Xu, Y. J. Li, L. Y. Wang, C. X. Liu, and T. Y. Chai, “Monitoring system for mineral processing equipment based on IoT and industrial cloud computing,” *Computer Integrated Manufacturing Systems*, vol. 23, no. 009, pp. 1972–1982, 2017.