Intelligent and Converged Networks ISSN 2708-6240
2022, 3(4): 325–339 DOI: 10.23919/ICN.2022.0026

IoT data cleaning techniques: A survey

Xiaoou Ding, Hongzhi Wang*, Genglong Li, Haoxuan Li, Yingze Li, and Yida Liu

Abstract: Data cleaning is considered as an effective approach of improving data quality in order to help practitioners and researchers be devoted to downstream analysis and decision-making without worrying about data trustworthiness. This paper provides a systematic summary of the two main stages of data cleaning for Internet of Things (IoT) data with time series characteristics, including error data detection and data repairing. In respect to error data detection techniques, it categorizes an overview of quantitative data error detection methods for detecting single-point errors, continuous errors, and multidimensional time series data errors and qualitative data error detection methods for detecting rule-violating errors. Besides, it provides a detailed description of error data repairing techniques, involving statistics-based repairing, rule-based repairing, and human-involved repairing. We review the strengths and the limitations of the current data cleaning techniques under IoT data applications and conclude with an outlook on the future of IoT data cleaning.

Key words: Internet of Things (IoT); data quality; data cleaning; error detection; data repairing

1 Introduction

The Internet of Things (IoT) is producing data at an unprecedented rate with the explosive growth of information industry. IoT data become an increasingly significant asset[1], due to its growing application demand in human’s activities. IoT data with time series characteristics are usually a series of data points that can include different types of measurement taken at regular intervals and are indexed in time order. The massive IoT data flowing from the physical world to the digital world will promote the realization of statistics analysis of big data, lead the development and transformation of relevant fields, and play a key role in IT industry in the future.

High-quality IoT data are vital to the activities such as business process management, decision support, demand analysis, service quality improvement, risk prediction, etc. The accuracy and reliability of IoT data are the basis of the full use of the advantage of big data. However, the erroneous and problematic features are thought to be inherently associated with IoT data[1], as summarized in Table 1. Low-quality data inevitably affect the quality of information extracted from it, and even lead to invalid and incorrect information. The pervasive IoT data quality issues not only have a high negative impact on the downstream application, but also cause severe loss of data[2]. In Table 2, we classify data quality issues in IoT data in four categories according to the presentation of them in the dataset. With regard to IoT data with time series characteristics, it is of great necessity to establish a complete and effective data cleaning roadmap to ensure its applicability.

The level of data quality represents the extent to which the data satisfy the quality metrics, and reflects the adaptability of data use and compliance with application requirements[3, 4]. Data quality issues are various and complex in IoT data. This paper provides a systematic summary of data cleaning techniques for IoT data. Our contributions in this paper include: (1) We present an overview of IoT data cleaning, including

• Xiaoou Ding, Hongzhi Wang, Haoxuan Li, Yingze Li, and Yida Liu are with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China. E-mail: {dingxiaoou, wangzh}@hit.edu.cn.
• Genglong Li is with the School of Mechatronics Engineering, Harbin Institute of Technology, Harbin 150001, China.
∗ To whom correspondence should be addressed.

Manuscript received: 2022-09-16; revised: 2022-11-18; accepted: 2022-12-17

Table 1 Factors contributing to IoT data quality issues.

Factor	Explanation
Variety of data formats	The data generated by various kinds of IoT devices include structured, semi-structured, and unstructured formats. When data are integrated, data quality issues may arise.
Inconsistent timestamps	When integrating data from multiple sources, timestamps may not be aligned. Besides, the generated data may be inconsistent or have null values because of the unaligned timestamps.
Faulty devices	The mechanical failure happens in sensors and other important monitoring devices in IoT applications owing to natural conditions, human damage, or using over an extended period.
Defective data transmission	Data collected by IoT devices are transmitted to back-end applications in the form of streams. Data quality problems would likely happen in this process.
Data retrieval redundancy	Data redundancy occurs when part of data exist in multiple places. Accidental data redundancy can lead to data inconsistency and data corruption, which increases the cost of data management.

Table 2 Types of IoT data quality issues.

Quality issue	Manifestation type of quality issues
Missing data	Some data instances contain missing or invalid values.
Data duplication	The same data instances occur multiple times in the dataset.
Data anomalies	The behaviors of some non-missing data instances deviate remarkably from those of others.
Data violation	Some pieces of data do not satisfy predefined quality rules or constraints.

introduction of data quality problems in IoT data applications. (2) We summarize the error data taxonomy, discuss fundamental strategies in both error detection and repairing tasks, and introduce the state-of-the-art data cleaning techniques. (3) We provide future work discussion about IoT data cleaning.

1.1 Data quality assessment and management methods

In order to provide a deeper insight into IoT data quality issues, we briefly review the development process of data quality assessment and management methods.

Since the end of the 20th century, the requirements for the concept of data quality focus on data accuracy[5]. And the requirement of data quality has been continuously extended in these years. Guo and Zhou[6] introduced the importance of data quality and metrics including data consistency, correctness, and completeness. It is an early work that introduced an in-depth analysis of data quality. Fan and Geerts[7] summarized important research issues such as data quality definition, measurement, analysis, and improvement. In the process of data quality assessment, data quality dimensions[8] describe the degree of compliance of data quality with established requirements in a particular aspect and provide an effective method for quantifying data quality. Li et al.[9−11] conducted an exhaustive study on the problem of big data availability, summarized data consistency, accuracy, completeness, currency, and entity homogeneity as the core dimensional indicators of data quality. They proposed several challenging problems such as data quality expression mechanism, automatic data error detection and repair, and low-quality data tolerance computing. Since the multiple important factors affecting data quality are not completely isolated, Ding et al.[11] proposed a comprehensive framework for data quality assessment based on multiple dimensions. They proposed a comprehensive summary of error forms in four important properties of data, and provided definition and quantification method of violation forms. In 2017, Cai and Zhu[12] summarized techniques of data quality management solutions in specific scenarios from the perspective of data assets and analyzed the technical challenges of research on big data quality.

With the increase of time series data collected from IoT environments, the problems of time series data quality have drawn more attention. On one hand, the rich data types and complex data structures increase the difficulty of data quality management; on the other hand, the large scale of data and higher practical requirements make it more difficult for data management methods to make reliable evaluation on the merits of data within a reasonable time period.

Reference [1] summarized data quality issues including missing data, inaccurate records, and uneven data clocks from multiple sources in IoT scenarios and introduced an error data cleaning model with outlier detection as the core. Nargesian et al.[4] proposed the challenges and opportunities to achieve data management for data lakes. Song et al.[13, 14] summarized how to assess and manage IoT data quality from three aspects including validity, completeness, and consistency.

1.2 Overview of data cleaning techniques

Data cleaning is an effective approach to solving the troublesome IoT data quality issues[9]. In this section, we first systematically summarize the development history of data cleaning, aiming to introduce the specific techniques applied to IoT data in the following chapters.

We consider data to be erroneous when they are identified to have a large distance from the true value they are considered to obtain[15, 16]. To get correct and valid "clean" data, it is necessary to detect and identify low-quality data and take appropriate methods to correct the errors. Data cleaning techniques are applied in the process of data quality improvement, which is known as a necessary step in the data preprocessing process[12, 16, 17]. It also provides the necessary foundation for subsequent data analysis.

The main objective of data cleaning is to design efficient methods to detect and repair errors in data as comprehensively and accurately as possible while minimizing the cost of manual operations[6, 18] . In recent years, quite a few survey papers[17, 19–21] have summarized data cleaning techniques from different perspectives and multiple types of scenario requirements. In 2018, Hao et al.[19] conducted a review of relational data cleaning techniques, which introduced current cleaning methods in four aspects: data missing, redundancy, conflict, and error. It analyzed the advantages, disadvantages, and applicability of advanced data cleaning techniques. In 2019, Ilyas and Chu[17] summarized the typical process of data cleaning, introducing the research difficulties and state-of-the-art techniques of data transformation, quality rule discovery, erroneous data detection, and erroneous data repair problems.

Error data detection and error data repair are the two key stages making up of data cleaning[17]. Erroneous data can be identified through error detection, and data quality requirements can be met through reasonable and effective repair of erroneous data. In this paper, we focus on the general techniques and the techniques designed for IoT data with time series characteristics. The rest of the paper is organized as follows. We introduce error data detection techniques in Section 2 and error data repair techniques in Section 3. We summarize challenges and future directions of the current research of IoT data cleaning techniques in Section 4. We conclude this survey in Section 5.

2 Error data detection

Error data detection is the first and key step in the data cleaning process. Only when the " dirty data" are accurately detected, people could repair them effectively. The research difficulties for error data detection include: how to identify the real error data from the identified violation data units and how to deal with the higher computational cost brought by the growth of the data size and the number of constraints. In 2016, Ref. [20] summarized that the error detection techniques can be either quantitative or qualitative based on how to define errors. Quantitative errors are defined as the data whose behavior remarkably deviates from that of other data, while qualitative errors are defined as data that violate the predefined constraints or patterns. Among them, quantitative errors mainly include abnormal values, outliers, etc., while qualitative errors mainly include duplicate records, pattern violations, and rule violations, etc.[22] We review the research progress of quantitative data error detection and qualitative data error detection methods which can be applied to IoT data in Section 2.1 and Section 2.2, respectively. Table 3 summarizes some typical error detection methods for both types of errors.

2.1 Quantitative data error detection

Quantitative data error detection aims to find outliers

Table 3 Summary of error data detection methods.

Type of the methods	Type of the errors	A brief summary of existing error data detection methods
Quantitative data error detection method	Single-point error	Earlier summaries: anomaly detection[23], temporal data outlier detection[24], and wireless sensor network data outlier detection[25].
	Continuous-type error	Time series data transformation representation methods: symbolic aggregation approximation (SAX) for subsequence mining methods[26]; discrete Fourier transform, singular value decomposition, segmented aggregation approximation[27], and inverse nearest neighbor based parameter-free detection method[28].
	Multidimensional time series data error	Combining statistical indicators and learning models method[29, 30].
Qualitative data error detection method	Rule violation error	Existing rule-based data quality expression mechanism: function dependencies (FD), conditional function dependencies (CFD)[7, 31], inclusion dependencies, metric dependencies, difference dependencies[19], sequential dependencies[32], temporal constraints[13, 33], and denial constraints (DC)[34]. Rule-based error detection method: DC-based[35−37].

with more significant deviations in data, usually using abnormal point detection techniques. Quantitative errors are one common form of errors in IoT data. Quantitative errors in IoT data with time series characteristics are mainly manifested as single-point errors, continuous errors, (contextual) content errors, etc.[21] In degrees of error occurrence, they can also be divided into large errors and small errors[2].

For single-point error detection techniques, earlier research work[38] defined generalized presentations such as additive outliers and innovative outliers. In 2014, Ref. [24] reviewed anomaly detection techniques for time series data and summarized existing techniques for a variety of data types such as time series data, data streams, and time-space data. In multiple fields such as anomaly detection and data cleaning, researchers have conducted extensive research on outlier detection to improve the correct detection rate, and proposed detection techniques including statistical model-based, classification-based, clustering-based, and nearest neighbor model based methods[23, 25] . We note that various detection techniques have their advantages and limitations in terms of parameter settings, computational efficiency, and sample requirements.

In addition to single-point error detection, researchers have also analyzed the “length” and “width” of error data and conducted studies on the detection of continuous errors and errors on multidimensional sequences[39]. In terms of continuous error detection, researchers proposed subsequence mining methods to identify anomalous patterns with a certain length. In 2005, the novel subsequence mining method of symbolic aggregate approximation (SAX) was proposed in Ref. [26], where sequences are transformed symbolically segment by segment to achieve classification and pattern discovery of subsequences. Other common methods for transforming the representation of time series data include the discrete Fourier transform, singular value decomposition, and segmented aggregation approximation, etc.[27] In 2020, Ref. [28] used concepts such as inverse nearest neighbor (INN) and proposed a parameter-free error data detection method by scoring the number size, variance, and correlation and reducing the cost of manual labeling to achieve the detection of single-point anomalies, collective anomalies, and change points in single-dimensional time series data.

Multidimensional time series data are the main type of IoT data. The error patterns of multidimensional IoT data are more diverse and complex, and it is necessary to consider not only the problem of single-point anomalies but also issues such as the correlation between variables. Reference [40] earlier initiated the study of correlation mining among numerical attributes on relational databases, and it achieved similarity calculation of numerical attribute combination by ranking metric. For multidimensional IoT data with temporal features, detection methods combining statistical metrics and learning models have been developed[29], and classification problem models represented by decision trees, support vector machines, and neural networks are used to add identification to dissimilar data. The method usually requires a certain size of anomalous instances as training data, and the main points of the method include sample analysis, feature extraction, anomaly scoring vector construction, feature selection, and anomaly pattern matching[30]. We note that effective samples and the feature extraction methods are common challenging factors to solve the anomaly detection of IoT data with classification models.

2.2 Qualitative data error detection

Qualitative data error detection aims to achieve high automation, accuracy, recall, and efficiency of detection methods, and the rule violation problem is one of the most important problems in qualitative data error detection. A research lineage has been formed in the direction of data availability expression mechanism, data quality determination, and erroneous data detection and repairing[9]. In the following, three points are introduced from the mechanism of rule-based data quality expression, discovery and exploration of data quality rules, and rule-based error detection. Most of the current qualitative data error detection methods are general methods which are not specially designed for IoT data, but they are effective to express some user-specified semantics and offer interpretation to the detected errors in IoT data. Therefore, they are important methods in IoT data cleaning as well.

(1) Rule-based data quality expression mechanism

A set of integrity constraints developed by domain experts or summarized by domain business processes is always regarded as data quality rules, which could accurately and comprehensively describe the semantics of the data domain[17]. Rule-based detection methods usually employ domain-specific knowledge to fully explore the patterns and relationships of data and use rules to improve data quality. Data that do not satisfy the given data quality rules are marked and judged as erroneous or violated data. Academia and industry have studied rule-based error data detection and repair methods for a long time, and the main points of the research locate in data quality constraints discovery and inference problem, the constraints coverage problem, and the violation detection problem[9].

Earlier, researchers have proposed function dependencies (FD), conditional function dependencies (CFD)[7, 31], and other semantic rules to achieve an effective representation of data consistency. In recent years, researchers have conducted a lot of theoretical and practical research on data quality rule expression and violation detection with conditional function dependencies. The established integrity constraints also include inclusion dependencies, metric dependencies, difference dependencies, etc.[19] Golab et al.[32] introduced the concept of sequential dependencies to describe the requirement of numerical differences between consecutive data points in sequential data, but did not consider the semantic representation of timestamps. Considering the quality requirements of data in terms of time attributes and addressing the problems of incomplete timestamps and lack of timely maintenance in databases, the concept of currency constraints with first-order logical statements[14, 33] was proposed to achieve currency measurement for attributes and tuple records in databases without relying on timestamps. Since 2013, researchers have proposed the denial constraints (DC)[34], a generic quantified first-order logical form of constraints, to improve the expressiveness of data quality. DC is both compatible with FD and CFD and it extends them to realize the comparison judgment of "greater than" and "less than" between attribute values. Since 2015, for IoT data with timestamps, Zhang et al. have proposed the concepts of speed constraints[2, 41] and acceleration constraints[42] for discovering single-point error data with different magnitudes of change on a single time series. In 2019, Pena et al.[43] proposed a definition and identification method for approximate denial constraints (ADC).

(2) Discovery and exploration of data quality rules

Data quality rule induction and preparation is costly in terms of time and labor[44], and in recent years, researchers have launched theoretical studies on automated or semi-automated rule mining, aiming to achieve discovery extraction of valuable rule patterns hidden in data[34, 43]. The data quality rule mining issues can be summarized as given a rule form and a set of clean data, discovering the minimum set of all correct and non-trivial rules in the data[17].

Currently, researchers have designed algorithms to implement the mining of conditional function dependencies and denial constraints. The early proposed function dependency mining methods include TANE[45] and FASTFD[46]. In 2015, Ref. [47] conducted a measurement comparison experiment on seven function-dependent mining methods. On the theoretical basis of mining methods for FD, researchers have continued to expand their research on CFD mining and proposed methods such as CTANE[48], CFDMiner[49], and FASTCFD[50].

Compared with the semantic system of function dependencies, the denial constraint expands the “equal to” relationship in function dependencies to the comparison relationship of “ greater than” and “less than”. The increase in the expressiveness of the denial constraint also leads to an increase in the computational space of the mining problem, which makes it more difficult to solve than the function-dependent mining problem. Typical denial constraint mining methods include FASTDC[34] and heuristic Hydra mining algorithm[51]. Since 2019, Refs. [43, 52] have proposed the concept of approximate denial constraints and their mining problem to avoid overfitting problems in rule mining by allowing a certain amount of relaxation in the degree of constraint satisfaction. For the approximate denial constraint, Livshits et al.[52] constructed the approximation function based on the denial constraint and proposed the mining algorithm ADCMiner.

(3) Rule-based error detection

Researchers have studied and proposed methods such as violation detection for denial constraints[35]. Chu et al. proposed a global data cleaning framework with denial constraints as the main form of data quality rules, which implemented error data detection method based on the conflict hypergraph model in 2013[36], and proposed Holoclean[37], a global data cleaning tool, in 2017. Holoclean is currently a common benchmark framework in the field of data cleaning research.

In the past decade, data quality rule expression with conditional function dependencies and denial constraints as the core has been established, and theory and technology system have been developed for rule semantic representation, rule discovery, violation detection, etc. Researchers improved and optimized the computational effectiveness and efficiency of the detection method from multiple perspectives, such as violation characterization, distance metric, and approximation rules discovery. However, it is easy to find that there remains limitations and shortcomings in the expressiveness of the current function-dependent rule system for complex data quality requirements in IoT data, and we will further summarize them in Section 4.

3 Error data repair

Compared to the detection step which focuses on the “identification” of the erroneous data, the key point of the data repairing step is how to reasonably and effectively repair the real erroneous data so that the repaired data meet the data quality requirements while avoiding misuse of the clean data to prevent excessive modification deviations to the original data[53]. The difficulties in the repair phase are: whether we can perform the repair on the real wrong data and whether we have to calculate the exact repair value. To reduce the workload of manual repair and improve the quality and repair efficiency of automated data repair, researchers have deeply studied the error data repair goals, repair methods, and repair models. Facing the key issues of repair efficiency, correctness and comprehensiveness of repair, statistics, rule-based, human involvement, and learning model based error data repair techniques have been developed as the main methods[19, 21]. In this section, we first review the relatively well-developed general error data repair techniques originally designed for relational data in Section 3.1, which can offer some references on or an inspiring look at the research of the currently insufficient IoT data repair techniques. Then, we review data repairing techniques especially for IoT data in Section 3.2.

3.1 General error repair techniques

General error repair techniques can be divided into automatic error data repair techniques and error data repair with human involved techniques. In regard to automatic error data repair, we focus on rule-based error data repair in this paper.

3.1.1 Automatic error data repair

For the theory of data repairing, Li et al.[9] proposed a computable theory for data error detection and repairing. In 2010, Fan et al.[54] proposed the concept of deterministic repairing and proved that minimizing the deterministic repairing region is an NP-complete problem. On this basis, Ref. [7] proposed a data cleaning framework UniClean. The data quality rules for cleaning are formulated by calculating the repair confidence as well as the entropy of the data uncertainty degree. The repair method with high confidence is preferentially selected to ensure the data repair effectiveness. In 2016, taking conditional function dependencies in relational data as the research object, Miao[55] investigated the computational complexity and evaluation algorithms for error data repairing, proposed a computational method for the effectiveness of the ruleset complexity on the computational complexity of the problem, and studied effective repair algorithms to achieve the removal or repair of errors. The above research results have laid the theoretical foundation for data repairing. Rule-based repair of erroneous data has two types of repair strategies[20]: (1) fully trusting the given set of data quality rules and only performing modification operations on the data and (2) not fully trusting the given set of rules and considering modification operations on both data and quality rules.

(1) The current research results are dominated by carrying out data-only modification repair strategies for a data quality rule that adopts the principle of minimizing modified data[35, 56] , defining a modification cost function, and designing a solution to satisfy the conditioned repair results. The modification cost here is usually a distance function or a cost function with similar properties to the distance function. The typical research is Holoclean[37], where the data units with a high probability of error occurrence and a high number of errors are modified iteratively according to the cost function by solving the minimum vertex coverage on the constructed conflict hypergraph.

(2) Considering the problem of incomplete and imprecise data quality rules, Chiang and Miller[57] first proposed to consider both modifying data as well as modifying rules in 2011. Given the inferior dataset and the function-dependent rule set, one needs to find the repaired dataset and the modified ruleset to minimize the repair cost such that the repaired dataset satisfies the modified rule set. Reference [57] used the description length of the rule to represent the editing distance of the data in collaboration with the rule and then proposed a new repair cost function to implement the operation of adding attribute items to functional dependencies. In 2013, Beskales et al.[58] continued to optimize the data-rule collaborative repair algorithm by introducing confidence refinement to modify the cost objective and achieve the addition of new attributes to the rules. In recent years, Song et al.[59] proposed a data and denial constraints collaborative repair method, which solves the problem of finding the set of denial constraints that minimizes the cost of modifying the dataset within a range of variable thresholds for a given constraint. The method addressed the problem of candidate rule search pruning and the sharing of repair information among different candidate rulesets to avoid overfitting and underfitting situations.

3.1.2 Human-involved error data repair

The complexity and uncertainty of data make it difficult for automated repair algorithms to repair erroneous data with 100% confidence and accuracy[60], and humans, represented by domain experts and users, are an important part of data cleaning, serving to improve the reliability and validity of methods in various aspects such as quality rule discovery and validation, erroneous data labeling, and repair pattern selection. In recent years, Fan et al.[61] proposed the concept of human-in-the-loop data preparation, which summarizes the manual participation methods and manual tasks in data preparation processes such as data extraction, annotation, integration, and cleaning. Compared to automated repair algorithms, manual participation repair has the advantages of high repair accuracy, reliability, and showing good results for domain-specific data repair, but it is blamed for its repair cost.

Currently, the two main strategies for manual participation in data restoration techniques include (1) crowdsourcing repairing and (2) human-machine interactive repairing.

(1) Crowdsourcing-based data repair is to transform repair problems into multiple "questions" to be answered and distributed to users through a crowdsourcing platform, making effective use of human knowledge and processing power at a low cost. In recent years, Li et al.[44] studied the basic theory and technology of crowdsourcing-based data cleaning from the aspects of repair quality, labor cost, and delay management calculation, and proposed difficult research problems such as improving the quality of crowdsourcing completion and reducing labor cost and response delay. Ye et al.[62] proposed a crowdsourcing-based method model for truth discovery, missing value filling, inconsistent data detection, and repair of relational data, which improves the repairing process of low-quality data. Reference [63] proposed KATARA, a knowledge base and crowdsourcing-based data cleaning system, which aims to compensate for the limitations of established integrity constraints and model-driven data cleaning in terms of effectiveness by accessing master data or asking for domain experts to resolve ambiguities in the data.

(2) Compared to crowdsourcing methods that focus on task assignment, human-machine interactive repair methods focus on the optimization of manual interaction mechanisms and the number of interactions, as well as the optimization of methods based on manual feedback. The three main human involvement approaches are quality rule validation, erroneous data determination, and repair strategy selection[64]. In 2011, Yakout et al.[65] proposed the classical GDR human-machine combined data cleaning model, which utilizes conditional function dependencies as quality rules, and the user gives feedback on the algorithm's decision results and cleaning strategy to correct incorrect decisions and retrain the learning model. In 2014, Volkovs et al.[66] proposed a continuous data cleaning model with elements of repair classifier, repair space search, and manual interaction. The model uses a classifier to determine whether to perform the repair on data, rules, or both. According to the result of the classifier, a search algorithm reduces the search space and recommends a set of candidate repairs to the user. Then, the user determines the applied repair from the candidates. The manual processing result is handed back to the repair classifier to achieve iterative incremental repair and solve the cold start problem of the repair model. In 2016, Ilyas[64] proposed a looped "cleaning-analysis" model based on continuous data cleaning and proposed manual participation methods in the maintenance of rule violation evidence sets, error data interpretation and traceability, and manual selection methods for repair strategies. In 2019, Rezig et al.[67] proposed a human-centered data cleaning model and defined four manual participation roles: data manager, user, validator, and domain expert, which regulates the form of manual participation and computational tasks in the data cleaning process. In 2019, in terms of the cost of human interaction, Ref. [68] proposed the problem of sequential feature explanations (SFEs) of error data. Whether an error instance needs to be returned to human judgment is determined by metrics such as information content and detection confidence. And the contribution of the returned error instance features to the error is ranked from the highest to lowest to achieve the goal of reducing the number of human involvement. In 2020, Ref. [69] proposed an error data cleaning model integrating crowdsourcing and active learning models. To rank the task processing results in terms of manual accuracy and response time, two user roles general human and domain expert were distinguished. The model solved the error data manual validation problem with the optimization objectives of minimizing expert participation cost and maximizing validation accuracy.

A summary of general error repair techniques is shown in Table 4. We note that there are still many key issues that have not been solved in the research of

Table 4 Summary of general error repair techniques.

Type	Repair strategy	A brief summary of existing general error repair techniques
Automatic rule based error data repair	(1) Modifying data only while trusting given data quality rules[54, 58, 60]	Cost-minimal error repair: define a modification cost function, and design a solution to satisfy the conditioned repair results. Holoclean[37] is a typical method.
	(2) Modifying both data and given data quality rules[59, 61, 62]	(1) Make use of the description length of rules to construct a unified model for repairing data and function dependencies on an equal footing[57]. (2) Introduce confidence refinement to modify the cost objective to add quality rule (FD) attributes[58]. (3) Solve the problem of finding the set of denial constraints that minimizes the cost of modifying the dataset within a range of variable thresholds for a given constraint, pruning candidate rule, and sharing repair information[59].
Human-involved error data repair	(1) Crowdsourcing-based[64–66]	(1) Crowdsourcing-based model[62] and (2) knowledge base and crowdsourcing-based data cleaning system—KATARA[63].
	(2) Human-machine collaborative[68–70]	(1) GDR model[65]; (2) continuous data cleaning model[66]; and (3) end-to-end human-centric data cleaning framework[67].

human-machine combined cleaning. In the face of the large amount of labor and time costs associated with manual participation, it is usually necessary to make a comprehensive assessment of efficiency, effectiveness, and cost before selecting a reasonable cleaning strategy.

3.2 Error repair techniques for IoT data

In this section, we review two kinds of repairing techniques which are widely applied in IoT data cleaning.

3.2.1 Statistics-based error repair

The maximum likelihood method is the core method for solving statistics-based repair problems by calculating the data distribution with probabilistic indicators to obtain the best repair results. In relational data cleaning, both the SCARED method[70] and the Holoclean cleaning framework[37] are classical likelihood based repair methods. The SCARED method achieves effective field value repair by probabilistic modeling of reliable attributes and attributes containing erroneous data. However, these methods are usually designed for categorical attributes and are difficult to apply directly to the problem of error repair in IoT data[2]. Maximum likelihood models are commonly used to repair missing and erroneous values of indoor radio frequency identification devices (RFID) data[21], and Zhang[2] proposed a maximum likelihood based cleaning method for single point small magnitude errors in single time series and designed an exact solution method as well as an approximation algorithm considering different objectives to achieve effective detection and repair of small magnitude errors.

3.2.2 Rule-based error repair

The rule-based constraint cleaning technique is not only widely used in traditional relational data, but also applicable to the cleaning problem of IoT data. However, the quality rule expression system for IoT data including time series data has not been fully constructed yet. In 2016, Ref. [71] focused on time series data quality management and proposed four types of normalized data quality rules from the perspective of "rows" and "columns" : single-entity (row) single-attribute (column), single-entity multi-attribute, multi-entity single-attribute, and multi-entity multi-attribute. It proposed a general model for time series data quality management. First, the rule constraints to be satisfied by high-quality data are defined inductively, and the high-quality data at moment t are noted as D^l_t; then, the test data D_t are compared with the high-quality data D^l_t, and the difference between the test data D_t and the ideal data D^l_t, i.e., $\mathrm{SD}(D_t) = \mathrm{Dist}(D_t, D^l_t)$, is measured using the statistical distortion metric proposed in Ref. [48] to find the data that violate the constraints. The violated data initially detected by constraints are also referred to as small errors[72]. It is important to note that small errors detected using constraints are not necessarily real errors and may be misclassified due to overly strict parameter values set for the constraints. Therefore, the violation data usually need to be further analyzed and interpreted before making decision on them.

At present, the available research results mainly focus on single-entity multi-attribute and multi-entity single-attribute, and the research on multi-entity multi-attribute rules is still insufficient. Both function dependencies and denial constraints are typical representatives of the single-entity multi-attribute type, but considering the time-window factor on time series data, the single-entity multi-attribute rule describes the quality requirements of single point-in-time data and fails to express the quality requirements of time-window data. Considering the regular representation of the time series, the problem of cleaning the time series data based on velocity constraints was proposed and solved in Refs. [41, 73] to achieve the repair of substantially erroneous data on a single sequence. Since only a single attribute variable is involved, the speed constraint based repairing method on a single sequence can be converted to a linear programming problem for solving, compared to the NP time complexity of conditional function-dependent and denial-constrained restoration algorithms, which is earlier research on rule-based data cleaning. On this basis, in 2018, Yin et al.[74] proposed data cleaning method combining variance constraints with speed constraints. In 2021, Gao et al.[75] proposed a repair method with multi-interval speed constraints, and Song et al.[42] proposed a repair method combining speed constraints and acceleration constraints, which improved the practicality of multi-entity single-attribute rules. Due to the limitations in semantic expressiveness, few repair methods based on multi-entity multi-attribute rules are currently carried out. In 2020, the concept of similarity rules was proposed in Ref. [76] and used to solve the vacancy value filling problem for time series data. Ding et al.[77, 78] proposed a correlation-based error detection method for multidimensional time series data, a framework for a multi-role error identification and diagnosis method for four types of data quality rule types, and a solution method based on a weighted set coverage. On this basis, in 2021, Liang et al.[79] conducted an example study on power plant IoT timing data to solve the key feature calculation problem of time series data errors using variance constraints, velocity constraints, and similarity rules.

The time series data repair techniques are summarized in Table 5. Compared to the research on data error detection, the research on IoT data error repair techniques is still inadequate. Among the current repair methods, smoothing repair methods based on statistical models are predominant. However, its repair accuracy is difficult to meet the practical requirements. Although rules applicable to IoT data, such as sequential dependencies and velocity constraints, have been proposed, the theory of rule-based IoT data repair has not yet been fully established compared with the application of conditional function dependencies and denial constraints to traditional relational data, and the theoretical approach to IoT data repair still needs to be improved.

4 Challenges and future directions

The current research on data quality assessment and management methods and error data detection and repair technologies is mainly oriented to traditional relational data and has formed theoretical and technical achievements with data quality expression mechanism, data quality determination, data error detection and

Table 5 Summary of error repair techniques for IoT data.

Type	Repair strategy	A brief summary of existing time series data error repair techniques
Statistics-based	Maximum likelihood	Repair missing and erroneous values of indoor RFID data[21]; repair single-point small magnitude errors for time series[2].
Rule-based	Single-entity multi-attribute	Repair methods based on velocity constraints[41, 73]; repair methods combining variance constraints with velocity constraints[74].
	Multi-entity single-attribute	Repair method with multi-interval velocity constraints[75]; repair method with a combination of velocity and acceleration constraints[42].
	Multi-entity multi-attribute	Solving vacancy value filling problems using similarity rules[76]; correlation-based methods for repairing multidimensional time series data[77, 78]; solving key feature calculation problems for time series data errors using variance constraints, velocity constraints, and similarity rules, etc.[79]

repair, and approximate calculation of poor quality tolerance as the main lines. However, facing the IoT data with time series characteristics, the accumulated basic theories and technical methods still have great limitations, and many key problems of time series data error detection and repair still need to be solved by breakthroughs.

It is noteworthy that although lots of classic general data cleaning techniques can be applied in IoT data to solve corresponding data quality problems to some degree, the unique features and practical requirements of IoT data with time-series characteristics compared to relational data bring more challenges to error data detection and error data repair. We suppose that these unsolved challenges will also lead the future research directions. Some of the challenges are listed as follows.

(1) Time-series data generated by IoT devices require quality rules to become much more semantically expressive data dependencies within the time interval and between the data in the same row or column. For the characteristics of time-series data, when constructing the quality expression mechanism of time-series data, it is necessary to fully consider the constraint characteristics of the combination of time windows and ranks, and realize the effective semantic representation of ordered time windows, multi-dimensional inter-sequence function operations, and other factors. This puts higher and more complex requirements on the expressive power of the semantic of the existing data quality rules. At the same time, the rich quality rule representations also bring multiple challenges in terms of computational cost and computational effectiveness to the time-series data quality rule discovery and mining problems.

(2) Error in time-series data generated by IoT devices is characterized by various forms, with prominent cumulative effects and strong correlation. The error records of time series data not only occur in scattered time points but also often occur in periods. Moreover, the error patterns of the time series data are complex and diverse, with the phenomenon of “continuous slow change” in the period and “ sudden change” at the moment. In addition, there are effects of aggregation and accumulation of errors. Therefore, the detection method is required to have a strong ability to identify the error data with different deviation degrees and patterns and to avoid the occurrence of missed detection and false detection. At the same time, the strong correlation between the values of data records of different periods and a large number of attributes on the time-series data leads to the strong concealment of errors, which poses a great challenge to the accuracy and computational efficiency of the error detection method.

(3) The solution space for the repair of errors in time-series data generated by IoT devices is huge, and the uncertainty of the repair effect is high. The characteristics of time-series data such as time-series records and continuous accumulation of data values lead to a multi-dimensional real number field as the solution space for repair. Compared with the repair of relational data, the size of the repair space of time-series data is often more enormous, which brings a greater challenge to maintain the high efficiency of repair calculation. In addition, it is difficult to establish an effective and reliable repair model for time-series data in terms of repair index selection and repair effect measurement, which makes the repair results of the automated algorithm differ significantly from the theoretical true values and be prone to over-fitting or under-fitting, making the repair effect unable to meet the expected demand.

5 Conclusion

This paper introduces the concepts and development history related to data quality assessment and management methods. The cleaning methods for IoT data are deeply explored. The techniques of error data detection and error data repair, which are particularly important in data cleaning, are highlighted. We summarize the categories of methods involved in quantitative data error detection and qualitative data error detection, summarize the research progress of general error repair techniques and IoT data error repair techniques, and explain the application areas and limitations of the reviewed methods. Furthermore, we propose several challenges of the current research and the future directions of IoT data cleaning techniques.

Acknowledgment

This work was supported by the National Key Research and Development Program of China (No. 2021YFB3300502), National Natural Science Foundation of China (NSFC) (Nos. 62202126 and 62232005), and Heilongjiang Postdoctoral Financial Assistance (No. LBH-Z21137).

References

[1] A. Karkouch, H. Mousannif, H. A. Moatassime, and T. Noël, Data quality in internet of things: A state-of-the-art survey, *J. Netw. Comput. Appl.*, vol. 73, pp. 57–81, 2016.

[2] A. Zhang, Research on time series data cleaning methods, (in Chinese), PhD dissertation, School of Software, Tsinghua University, Beijing, China, 2018.

[3] K. Yue, *Data Engineering: Processing, Analysis and Services*, (in Chinese). Beijing, China: Tsinghua University Press, 2013.

[4] F. Nargesian, E. Zhu, R. J. Miller, K. Q. Pu, and P. C. Arocena, Data lake management: Challenges and opportunities, *Proc. VLDB Endow.*, vol. 12, no. 12, pp. 1986–1989, 2019.

[5] R. Y. Wang and D. M. Strong, Beyond accuracy: What data quality means to data consumers, *J. Manag. Inf. Syst.*, vol. 12, no. 4, pp. 5–33, 1996.

[6] Z. Guo and A. Zhou, Research on data quality and data cleaning: A survey, (in Chinese), *Journal of Software*, vol. 13, no. 11, pp. 2076–2082, 2002.

[7] W. Fan and F. Geerts, *Foundations of Data Quality Management*. San Rafael, CA, USA: Morgan & Claypool Publishers, 2012.

[8] F. Sidi, P. H. S. Panahy, L. S. Affendey, M. A. Jabar, H. Ibrahim, and A. Mustapha, Data quality: A survey of data quality dimensions, in *Proc. 2012 International Conference on Information Retrieval & Knowledge Management*, Kuala Lumpur, Malaysia, 2012, pp. 300–304.

[9] J. Li, H. Wang, and H. Gao, State-of-the-art of research on big data usability, (in Chinese), *Journal of Software*, vol. 27, no. 7, pp. 1605–1625, 2016.

[10] J. Li and X. Liu, An important aspect of big data: Data usability, (in Chinese), *Journal of Computer Research and Development*, vol. 50, no. 6, pp. 1147–1162, 2013.

[11] X. Ding, H. Wang, X. Zhang, J. Li, and H. Gao, Association relationships study of multi-dimensional data quality, (in Chinese), *Journal of Software*, vol. 27, no. 7, pp. 1626–1644, 2016.

[12] L. Cai and Y. Zhu, *Big Data Quality*, (in Chinese). Shanghai, China: Shanghai Science and Technology Press, 2017.

[13] S. Song and A. Zhang, IoT data quality, in *Proc. 29th ACM International Conference on Information & Knowledge Management*, Virtual event, Ireland, 2020, pp. 3517–3518.

[14] Z. Liu, Y. Zhang, R. Huang, Z. Chen, S. Song, and J. Wang, EXPERIENCE: Algorithms and case study for explaining repairs with uniform profiles over IoT data, *J. Data Inf. Qual.*, vol. 13, no. 3, pp. 1–17, 2021.

[15] W. Y. Kim, B. -J. Choi, E. K. Hong, S. -K. Kim, and D. Lee, A taxonomy of dirty data, *Data Mining and Knowledge Discovery*, vol. 7, no. 1, pp. 81–99, 2003.

[16] A. Chalamalla, I. F. Ilyas, M. Ouzzani, and P. Papotti, Descriptive and prescriptive data cleaning, in *Proc. 2014 ACM SIGMOD International Conference on Management of Data*, Snowbird, UT, USA, 2014, pp. 445–456.

[17] I. F. Ilyas and X. Chu, *Data Cleaning*. New York, NY, USA: Association for Computing Machinery, 2019.

[18] M. M. Lahijani, Semi-supervised data cleaning, PhD dissertation, School of Electrical Engineering and Informatics, Technical University of Berlin, Berlin, Germany, 2020.

[19] S. Hao, G. Li, J. Feng, and N. Wang, Survey of structured data cleaning methods, (in Chinese), *Journal of Tsinghua University* (*Science and Technology*), vol. 58, no. 12, pp. 1037–1050, 2018.

[20] X. Chu, I. F. Ilyas, S. Krishnan, and J. Wang, Data cleaning: Overview and emerging challenges, in *Proc. 2016 International Conference on Management of Data*, San Francisco, CA, USA, 2016, pp. 2201–2206.

[21] X. Wang and C. Wang, Time series data cleaning: A survey, *IEEE Access*, vol. 8, pp. 1866–1881, 2019.

[22] Z. Abedjan, X. Chu, D. Deng, R. C. Fernandez, I. F. Ilyas, M. Ouzzani, P. Papotti, M. Stonebraker, and N. Tang, Detecting data errors: Where are we and what needs to be done, *Proc. VLDB Endow.*, vol. 9, no. 12, pp. 993–1004, 2016.

[23] V. Chandola, A. Banerjee, and V. Kumar, Anomaly detection: A survey, *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, 2009.

[24] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, Outlier detection for temporal data: A survey, *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 9, pp. 2250–2267, 2014.

[25] Y. Zhang, N. Meratnia, and P. Havinga, Outlier detection techniques for wireless sensor networks: A survey, *IEEE Communications Surveys & Tutorials* , vol. 12, no. 2, pp. 159–170, 2010.

[26] E. Keogh, J. Lin, and A. Fu, HOT SAX: Efficiently finding the most unusual time series subsequence, in *Proc. Fifth IEEE International Conference on Data Mining* (*ICDM '05*), Houston, TX, USA, 2005, pp. 226–233.

[27] U. Rebbapragada, P. Protopapas, C. E. Brodley, and C. R.

Alcock, Finding anomalous periodic time series, *Mach. Learn.*, vol. 74, no. 3, pp. 281–313, 2009.

[28] K. -H. Le and P. Papotti, User-driven error detection for time series with events, in *Proc. 2020 IEEE 36th International Conference on Data Engineering* (*ICDE*), Dallas, TX, USA, 2020, pp. 745–757.

[29] S. M. Erfani, S. Rajasegarar, S. Karunasekera, and C. Leckie, High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning, *Pattern Recognit.*, vol. 58, pp. 121–134, 2016.

[30] R. Fujimaki, T. Nakata, H. Tsukahara, A. Sato, and K. Yamanishi, Mining abnormal patterns from heterogeneous time-series with irrelevant features for fault event detection, *Stat. Anal. Data Min.*, vol. 2, no. 1, pp. 1–17, 2009.

[31] I. F. Ilyas and X. Chu, Trends in cleaning relational data: Consistency and deduplication, *Foundations and Trends in Databases*, vol. 5, no. 4, pp. 281–393, 2015.

[32] L. Golab, H. J. Karloff, F. Korn, A. Saha, and D. Srivastava, Sequential dependencies, *Proc. VLDB Endow.*, vol. 2, no. 1, pp. 574–585, 2009.

[33] W. Fan, F. Geerts, N. Tang, and W. Yu, Conflict resolution with data currency and consistency, *J. Data Inf. Qual.*, vol. 5, nos. 1&2, pp. 1–37, 2014.

[34] X. Chu, I. F. Ilyas, and P. Papotti, Discovering denial constraints, *Proc. VLDB Endow.*, vol. 6, no. 13, pp. 1498–1509, 2013.

[35] P. Bohannon, W. Fan, M. Flaster, and R. Rastogi, A cost-based model and effective heuristic for repairing constraints by value modification, in *Proc. 2005 ACM SIGMOD International Conference on Management of Data* (*SIGMOD '05*), Baltimore, MD, USA, 2005, pp. 143–154.

[36] X. Chu, I. F. Ilyas, and P. Papotti, Holistic data cleaning: Putting violations into context, in *Proc. 2013 IEEE 29th International Conference on Data Engineering* (*ICDE*), Brisbane, Australia, 2013, pp. 458–469.

[37] T. Rekatsinas, X. Chu, I. F. Ilyas, and C. Ré, HoloClean: Holistic data repairs with probabilistic inference, *Proc. VLDB Endow.*, vol. 10, no. 11, pp. 1190–1201, 2017.

[38] R. S. Tsay, Outliers, level shifts, and variance changes in time series, *Journal of Forecasting*, vol. 7, no. 1, pp. 1–20, 1988.

[39] F. Moerchen, Algorithms for time series knowledge mining, in *Proc. 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD '06*), Philadelphia, PA, USA, 2006, pp. 668–673.

[40] T. Calders, B. Goethals, and S. Jaroszewicz, Mining rank-correlated sets of numerical attributes, in *Proc. 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD '06*), Philadelphia, PA, USA, 2006, pp. 96–105.

[41] S. Song, A. Zhang, J. Wang, and P. S. Yu, SCREEN: Stream data cleaning under speed constraints, in *Proc. 2015 ACM SIGMOD International Conference on Management of Data* (*SIGMOD '15*), Melbourne, Australia, 2015, pp. 827–841.

[42] S. Song, F. Gao, A. Zhang, J. Wang, and P. S. Yu, Stream data cleaning under speed and acceleration constraints, *ACM Trans. Database Syst.*, vol. 46, no. 3, pp. 1–44, 2021.

[43] E. H. M. Pena, E. C. D. Almeida, and F. Naumann, Discovery of approximate (and exact) denial constraints, *Proc. VLDB Endow.*, vol. 13, no. 3, pp. 266–278, 2019.

[44] G. Li, Y. Zheng, J. Fan, J. Wang, and R. Cheng, Crowdsourced data management: Overview and challenges, in *Proc. 2017 ACM International Conference on Management of Data* (*SIGMOD '17*), Chicago, IL, USA, 2017, pp. 1711–1716.

[45] Y. Huhtala, J. Kärkkäinen, P. Porkka, and H. Toivonen, Tane: An efficient algorithm for discovering functional and approximate dependencies, *Comput. J.*, vol. 42, no. 2, pp. 100–111, 1999.

[46] C. Wyss, C. Giannella, and E. Robertson, FastFDs: A heuristic-driven, depth-first algorithm for mining functional dependencies from relation instances extended abstract, in *Proc. Third Int. Conf. Data Warehousing Knowl. Discovery*, Munich, Germany, 2001, pp. 101–110.

[47] T. Papenbrock, J. Ehrlich, J. Marten, T. Neubert, J. -P. Rudolph, M. Schönberg, J. Zwiener, and F. Naumann, Functional dependency discovery: An experimental evaluation of seven algorithms, *Proc. VLDB Endow.*, vol. 8, no. 10, pp. 1082–1093, 2015.

[48] F. Chiang and R. J. Miller, Discovering data quality rules, *Proc. VLDB Endow.*, vol. 1, no. 1, pp. 1166–1177, 2008.

[49] W. Fan, F. Geerts, J. Li, and M. Xiong, Discovering conditional functional dependencies, *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 5, pp. 683–698, 2011.

[50] L. Golab, H. J. Karloff, F. Korn, D. Srivastava, and B. Yu, On generating near-optimal tableaux for conditional functional dependencies, *Proc. VLDB Endow.*, vol. 1, no. 1, pp. 376–390, 2008.

[51] T. Bleifuß, S. Kruse, and F. Naumann, Efficient denial constraint discovery with hydra, *Proc. VLDB Endow.*, vol. 11, no. 3, pp. 311–323, 2017.

[52] E. Livshits, A. Heidari, I. F. Ilyas, and B. Kimelfeld, Approximate denial constraints, *Proc. VLDB Endow.*, vol. 13, no. 10, pp. 1682–1695, 2020.

[53] T. Dasu and J. M. Loh, Statistical distortion: Consequences of data cleaning, *Proc. VLDB Endow.*, vol. 5, no. 11, pp. 1674–1683, 2012.

[54] W. Fan, J. Li, S. Ma, N. Tang, and W. Yu, Towards certain fixes with editing rules and master data, *Proc.*

VLDB Endow., vol. 3, nos. 1&2, pp. 173–184, 2010.

[55] D. Miao, Research on computational complexity theory and algorithms for data consistency, (in Chinese), PhD dissertation, School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, 2016.

[56] G. Beskales, I. F. Ilyas, and L. Golab, Sampling the repairs of functional dependency violations under hard constraints, *Proc. VLDB Endow.*, vol. 3, nos. 1&2, pp. 197–207, 2010.

[57] F. Chiang and R. J. Miller, A unified model for data and constraint repair, in *Proc. 2011 IEEE 27th International Conference on Data Engineering*, Hannover, Germany, 2011, pp. 446–457.

[58] G. Beskales, I. F. Ilyas, L. Golab, and A. Galiullin, On the relative trust between inconsistent data and inaccurate constraints, in *Proc. 2013 IEEE 29th International Conference on Data Engineering* (*ICDE*), Brisbane, Australia, 2013, pp. 541–552.

[59] S. Song, H. Zhu, and J. Wang, Constraint-variance tolerant data repairing, in *Proc. 2016 International Conference on Management of Data* (*SIGMOD '16*), San Francisco, CA, USA, 2016, pp. 877–892.

[60] S. Hao, Research on the key technology of cleaning structured data, PhD dissertation, Department of Computer Science and Technology, Tsinghua University, Beijing, China, 2018.

[61] J. Fan, Y. Chen, and X. Du, Progress on human-in-the-loop data preparation, (in Chinese), *Big Data*, vol. 5, no. 6, pp. 3–18, 2019.

[62] C. Ye, H. Wang, H. Gao, and J. Li, Active learning approach for crowdsourcing-enhanced data cleaning, (in Chinese), *Journal of Software*, vol. 31, no. 4, pp. 1162–1172, 2020.

[63] X. Chu, J. Morcos, I. F. Ilyas, M. Ouzzani, P. Papotti, N. Tang, and Y. Ye, KATARA: Reliable data cleaning with knowledge bases and crowdsourcing, *Proc. VLDB Endow.*, vol. 8, no. 12, pp. 1952–1955, 2015.

[64] I. F. Ilyas, Effective data cleaning with continuous evaluation, *IEEE Data Eng. Bull.*, vol. 39, no. 2, pp. 38–46, 2016.

[65] M. Yakout, A. K. Elmagarmid, J. Neville, M. Ouzzani, and I. F. Ilyas, Guided data repair, *Proc. VLDB Endow.*, vol. 4, no. 5, pp. 279–289, 2011.

[66] M. Volkovs, F. Chiang, J. Szlichta, and R. J. Miller, Continuous data cleaning, in *Proc. 2014 IEEE 30th International Conference on Data Engineering*, Chicago, IL, USA, 2014, pp. 244–255.

[67] E. K. Rezig, M. Ouzzani, A. K. Elmagarmid, W. G. Aref, and M. Stonebraker, Towards an end-to-end human-centric data cleaning framework, in *Proc. Workshop on Human-In-the-Loop Data Analytics* (*HILDA '19*), Amsterdam, the Netherlands, 2019, pp. 1–7.

[68] M. A. Siddiqui, A. Fern, T. G. Dietterich, W. -K. Wong, Sequential feature explanations for anomaly detection, *ACM Trans. Knowl. Discov. Data*, vol. 13, no. 1, pp. 1–22, 2019.

[69] H. Zhang, C. Chai, A. Doan, P. Koutris, and E. Arcaute, Manually detecting errors for data cleaning using adaptive crowdsourcing strategies, in *Proc. 23rd International Conference on Extending Database Technology* (*EDBT*), Copenhagen, Denmark, 2020, pp. 311–322.

[70] M. Yakout, L. Berti-Équille, and A. K. Elmagarmid, Don't be scared: Use scalable automatic repairing with maximal likelihood and bounded changes, in *Proc. 2013 ACM SIGMOD International Conference on Management of Data* (*SIGMOD '13*), New York, NY, USA, 2013, pp. 553–564.

[71] T. Dasu, R. Duan, and D. Srivastava, Data quality for temporal streams, *IEEE Data Eng. Bull.*, vol. 39, no. 2, pp. 78–92, 2016.

[72] L. Berti-Équille, T. Dasu, and D. Srivastava, Discovery of complex glitch patterns: A novel approach to quantitative data cleaning, in *Proc 2011 IEEE 27th International Conference on Data Engineering*, Hannover, Germany, 2011, pp. 733–744.

[73] A. Zhang, S. Song, J. Wang, and P. S. Yu, Time series data cleaning: From anomaly detection to anomaly repairing, *Proc. VLDB Endow.*, vol. 10, no. 10, pp. 1046–1057, 2017.

[74] W. Yin, T. Yue, H. Wang, Y. Huang, and Y. Li, Time series cleaning under variance constraints, in *Proc. 2018 Int. Workshops Database Syst. Adv. Appl.*: *BDMS*, *BDQM*, *GDMA*, *and SeCoP*, Gold Coast, Australia, 2018, pp. 108–113.

[75] F. Gao, S. Song, and J. Wang, Time-series data cleaning under multi-speed constraints, (in Chinese), *Journal of Software*, vol. 32, no. 3, pp. 689–711, 2021.

[76] S. Song, Y. Sun, A. Zhang, L. Chen, and J. Wang, Enriching data imputation under similarity rule constraints, *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 2, pp. 275–287, 2020.

[77] X. Ding, S. Yu, M. Wang, H. Wang, H. Gao, and D. Yang, Anomaly detection on industrial time series based on correlation analysis, *Journal of Software*, vol. 31, no. 3, pp. 726–747, 2020.

[78] Z. Li, X. Ding, and H. Wang, An effective constraint-based anomaly detection approach on multivariate time series, in *Proc. 4th International Joint Conference, APWeb-WAIM*, Tianjin, China, 2020, pp. 61–69.

[79] Z. Liang, H. Wang, X. Ding, and T. Mu, Industrial time series determinative anomaly detection based on constraint hypergraph, *Knowl. Based Syst.*, vol. 233, p. 107548, 2021.

Xiaoou Ding received the PhD degree from Harbin Institute of Technology in 2021. She is currently an assistant professor in the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. Her research work covers data cleaning, temporal data quality management, and temporal data mining. She has published more than 14 academic papers in various international conferences and journals in database community, including TKDE, VLDB, ICDE, CIKM, and APWeb-WAIM, and published 3 papers in the top Chinese journals. She has won the excellent academic paper award from China Association for Science and Technology (CAST) in 2020.

Hongzhi Wang received the PhD degree in computer science from Harbin Institute of Technology, Harbin, China in 2008. From 2008 to 2010, he was an assistant professor with Harbin Institute of Technology, China. From 2010 to 2015, he was an associate professor. Since 2015, he has been a professor and doctoral supervisor of the School of Computer Science and Technology. His research interests include big data management, data quality, graph data management, and web data management. He has published more than 100 papers in refereed journals and conferences. He is a recipient of the Outstanding Dissertation Award of CCF, Microsoft fellow, Chinese excellent database engineer, and IBM PhD fellowship.

Genglong Li is currently pursuing the BS degree at the School of Mechatronics Engineering, Harbin Institute of Technology. His current research interests include time series data cleaning and data quality.

Haoxuan Li is currently pursuing the BS degree at the School of Computer Science and Technology, Harbin Institute of Technology. His research interests include data cleaning, computer vision, and machine learning.

Yingze Li is currently pursuing the BS degree in data science and big data technology at Harbin Institute of Technology. His research interests include time series data quality management and database.

Yida Liu received the BEng degree from Harbin Institute of Technology, Harbin, China in 2022. He is currently pursuing the PhD degree in computer science and technology at Harbin Institute of Technology. His research interests include data cleaning, anomaly detection, and relaxed functional dependencies.