Conflict resolution and missing completion in the fusion of domain ontology in cyber-physical systems

Xiangfei Yan[1,2] and *Liwei* Zheng[1,2,*]

[1]Software Engineering Research Center of BISTU
[2]School of Computer Science, Beijing information Science and Technology University (BISTU), Beijing, China

Abstract. CPS integrates information services, human resource services, and physical equipment services, and always be supported by the ontologies in multiple domains. Due to the inconsistency of ontologies from different domains, the fusion of domain ontologies may have conflicts and deficiencies. Therefore, this paper provides a method for the conflict resolution and missing completion in the fusion of domain.

Keywords: Cyber-physical systems, Trustable functional dependency, Conflict resolution, Missing Completion.

1 Introduction

CPS (Cyber-Physical Systems) refers to an engineering system composed of a group of highly integrated physical and software components. The ontology of a certain field refers to the organic combination of information and relations in the field. CPS integrates information services, human resource services, and physical equipment services, and the information services may involve the collection of data or the integration of ontology in multiple domains. Due to the inconsistency of ontology in different domains, the fusion result of ontology may have conflicts and deficiencies.[1]

The sources of ontology are diverse. With the development of data collection technology, the scale of data to be managed has also grown rapidly. The forms and sources of data generation are diversified. The form of data organization is no longer single, and the relationship between data is no longer a pure one-to-one relationship in the past. It is a complex, multi-dimensional relationship. However, the increase in the amount of data does not mean the improvement of data quality. The increase in the amount of data is accompanied by an increase in the probability of data quality problems. New data relationships will also lead to new data quality problems. However, data quality issues may cause different problems in different fields. The same point is that they may have very serious consequences. For example, it may cause huge property losses in the financial field, and the medical field may lead to improper treatment of patients' conditions and even affect

* Corresponding author: zlw@bistu.edu.cn

the doctor-patient relationship. Therefore, it is very necessary to analyse the data quality problems in the context of big data and explore reasonable solutions. [2]

The advent of the data age means that the previous methods of processing data will no longer be fully applicable, and the role of data quality research in big data research has become more prominent. Generally, what people call high-quality data refers to the availability of data legibility, completeness, etc. For inferior data, in a narrow sense, it usually means that the data has one or more problems in terms of missing, redundancy, conflict, etc. In a broad sense, the credibility, relevance and other aspects of the data should also be considered. The process of turning inferior data into high-quality data through a series of methods and means is usually called data repair. For different data problems, the methods of data repair are also different. For data quality problems in relational databases, it is a feasible method to repair the data quality problem by using the method of functional dependence combined with conditional constraints. For data quality problems in non-relational databases with completely different data organization structures, relying solely on traditional functional dependence combined with conditional constraints is not enough to deal with, because different types of data, such as value pairs, graphs, documents, etc, are involved.

In terms of the expression mechanism of data availability, the advantages of relational databases are mainly reflected in the good maintenance of data consistency. Due to the high degree of data coupling, most of the data quality problems can be solved through traditional functional dependence and conditional constraints. For non-relational databases, data quality problems are not only reflected in data loss, redundancy, conflicts, etc., but also may involve data security or relevance issues [3]. For this type of data quality problem, the traditional method of restoring data quality with functional dependence and conditional constraints is often impossible to start. Therefore, for data quality problems that appear in data, this article believes that a research idea that can be tried is to redefine a expression form and processing mechanism of functional dependence and conditional constraint. Functional dependence can be used to express the mapping relationship between two data sets. Let R denote a relationship, A and B are two data sets containing this relationship, and use $A \to B$ to denote a functional dependence, if any two tuples in have equal values in set A, it can be inferred that their values in set B are also equal.

Based on the above theoretical basis, it can be concluded that the basic idea of using functional dependence to repair data quality problems is: define a functional dependence set on R, and if all functional dependencies in the set can be satisfied, then it is determined that the data quality requirements of R meet the standard. On the contrary, it is considered that there is a data quality problem [1]. Therefore, based on the traditional data restoration method for relational data using functional dependence as a means, this paper proposes a method to use dependency theory to repair the relationship quality problem of graph data.

In addition, the concept of graph involved in this article refers to the data stored in the graph structure, not the image data[4].The database used to store graph data is called the graph database. Graph is a type of non-relational database (NoSQL Database), usually used to process a large amount of graph-based data, suitable for the rapid growth of scientific data in recent years and data with similar structures such as social networks[5].Up to now, graph databases have been widely used in various fields. For example, processing a large number of configuration files generated by nodes in a network telemetry system. If these files are stored in a relational database, they will take up a lot of disk space and reduce efficiency; and if stored in a graph database, it is very convenient to process[6].For another example, map construction information can be extracted and inferred to achieve correlation analysis of information through the acquired quality information, and the results can be stored in the map database[7]. The article proposes the use of graph refinement methods for data pre-processing, and its purpose is to facilitate subsequent operations.

The organization structure of this paper is: Section 2 gives related work and proposes problems and research directions. Section 3 gives relevant definitions. Section 4 analyses examples and gives a repair model. Section 5 Summarize the work of this article.

2 Related work

Li Jianzhong[8] et al. summarized the research progress in various fields of big data availability in the article *Research Progress in Big Data Usability*, and summarized and looked forward to the future research directions. In the paper *Based on Functional Dependence and Conditions Constrained Data Repair Method*, Jin Cheqing and others proposed an algorithm that combines functional dependence and conditional constraints to repair data for relational databases and designed corresponding experiments[4].In the article, they first analysis the traditional use of functional dependence data repair strategy, and divided into two main strategies: one is to directly delete records that do not meet the requirements; the other is to not add/delete any records, but only modify certain fields. Through analysis, it is found that although the function dependence is very important and effective, but there are still some important constraints (hard constraints, quantity-related constraints, equivalent constraints, non-equivalent constraints, etc.) that cannot be described by functional dependence. Therefore, a data repair method combination functional dependence with conditional constraints is proposed. This method solves the shortcomings of only relying on functional dependence to repair data, but its own shortcoming is that the type of data that this method can handle is limited to relational data.

Hamza et al.[9] aimed at the problem of missing data in the Internet of Medical Things(IoMT) and proposed that a dynamic layer recurrent neural network(Dynamic L-RNN) can be constructed to predict missing data. The core idea is to use complete data set for deep learning, the trained model is used to predict the missing data in the incomplete data set, and finally achieve the purpose of repairing the missing data. The idea of this method is applicable to most of the data missing problems of relational or non-relational data. The disadvantage is that only the problem of missing data is analysed and processed, and other traditional data quality problems (such as data conflicts, etc.) are not involved.

M.TALHAa[10] et al. put forward the question of how to weigh data quality and data security in the big data environment, and analysed the conflicts and challenges that may arise. The author creatively put forward the view that data security may become an obstacle to data quality(vice versa), which may be ignored by most scholar. The conflict between the two systems makes the complexity more prominent. In the context of big data, flexible read and write permissions are necessary to implement a data quality management system, but this condition will leave data security risks, because this permission may be maliciously used by some people for illegal gains. For this reason, the author suggests that a fine-grained access control model can be implemented or extended to avoid conflicts between data security issues and data quality issues, available models include TBAC (Task Based Access Control), RBAC (Role Based Access Control), etc.

Danilo Ardagna[11] and others proposed a data quality service for the context-sensitive data quality evaluation problem in big data, which can evaluate the data quality of large data sets through parallel computing, and it can choose the amount of data for analysis on the basis of time and resource constraints. They mentioned that in the face of heterogeneous source processing, an adaptive method is required, which can trigger an appropriate quality assessment method based on the data type and context. The authors also considered that in some cases, due to performance and time constraints, it is impossible to evaluate the quality of the entire data set. Therefore, they proposed that only a part of the data set should be evaluated. Data quality evaluation, and by introducing credibility as the reliability metric of the quality evaluation program to measure the resulting loss of accuracy. From the result

point of view, this method effectively improves the efficiency of data quality evaluation, but it inevitably produces a loss of accuracy, although they make up for this loss to a certain extent by introducing credibility.

Maryam[12] and others proposed to use a structured learning theory combined with a data quality framework to explore the impact of processing big data on the quality of company decision-making and the mediating role of data quality and data diagnostics in this relationship, as well as by exploring big data The impact of utilization in terms of data quality and data diagnostics to improve the quality of corporate decision-making and revenue generation. The author found that although there are many related theoretical studies, there is no empirical study to explore the impact of big data utilization on the quality of corporate decision-making. Therefore, the author analysed the data of more than 130 companies, and put forward ideas for the impact of big data on data quality, data diagnostics, etc., and finally verified the ideas through quantitative experiments.

Fan Wenfei[13][14] et al. proposed a type of consistency constraint called conditional function dependency, which captures data consistency by enforcing semantic related value binding. He proposed a model, through partial Time sequence and time constraints are used to specify the timeliness of data, and the timeliness of data is strengthened through invariable conditional function dependence[15][16]. Carlo[17] and others have combined the detection and repair of data timeliness errors. Time-effect function dependence and approximate function dependence, the time-effect approximate function dependence is proposed, and its basic definition and some related data mining techniques are given.

Data sampling technology is usually used to improve the performance of the learner when the data is unbalanced. If the data quality is too low or the training data set is too small, the training results will be more unreliable. Jason Van Hulse conducted a comprehensive study on the characteristics of different training data sets, and the results showed that there are multiple indicators that affect the training results. Therefore, in the process of data set analysis, multiple indicators need to be considered[18].

Data dependence and fuzzy data dependence play a great role in maintaining data consistency and preventing data redundancy. P.C.SAXENA et al. standardized the concepts in Type 2 fuzzy relational database and defined new fuzzy functional dependencies[19].

Tu Feifei[20] and others summarized the data quality problems in software development support tools such as the problem tracking system and version control system, and summarized 9 data quality problems, and further proposed the use of redundant data to make corrections. And the method of mining user behaviour patterns to modify. The author analyses data quality problems from the three stages of data generation, data collection and data use, including: problem reports in the data generation stage, incorrect creation time and version control data Time issues; incomplete data crawling in the data collection stage and incomplete data issues caused by security and privacy; future data leakage issues in the data use stage, email address issues, and issues about authors and submitters in version control data.

From the above analysis, it can be seen that the method of functional dependence combined with conditional constraints has traditionally been mainly used to repair relational data, while data storage methods have undergone tremendous changes in the context of big data[21], and non-relational storage has gradually become the mainstream. The old conditional function dependence theory is no longer fully applicable, so this article attempts to study the new processing mechanism of function dependence in the context of big data combined with conditional constraints to solve the problem of graph data quality.

3 Basic concept definition

3.1 Domain ontology

The ontology is usually represented by a directed graph G. Suppose there is a data node set V, a label set P, and an attribute name set $Attrs$; for an attribute a in the attribute name set $Attrs$, its domain can be denoted as $Dom(a)$.The definition of knowledge ontology is given below[22].

Definition 1. A domain ontology G with data information can usually be represented by a two-tuple (V, R, ζ_V, ζ_R), where:

V is a finite set of data vertices,

$\zeta_V = \{\zeta_V^a \mid a \in Attrs \& \zeta_V^a : V \to Dom(a)\}$ is a function to specify attribute values for nodes,

$\zeta_R = \{\zeta_r^a \mid a \in Attrs \& r \in R \& \zeta_r^a : Dom(r) \to Dom(a)\}$.

Suppose that a directed graph G contains m nodes and n edges, where $m > 0$ and $n \geq 0$, $V(G) = (V_1, V_2, \ldots, V_m)$ and $E(G) = (E_1, E_2, \ldots, E_n)$ respectively represent the set of points and edges of the directed graph, and $P = Attrs(V) = (a_1, a_2, \ldots)$ represents the attribute set of the node. The attribute set V of a certain node V specific attribute A is represented as $a_x[A]$; the storage method of the d attribute is in the form of a key-value pair. Use $key(a_1)$ to represent the key of attribute a_1, use $value(a_1)$ to represent the value of attribute a_1, and $key[V]$, $value[V]$ respectively represent the key of node V attribute set and value set. $R = (R_1, R_2, \ldots)$ means the set of relations between any two nodes, $R_1(V_1, V_2)$ means that the two nodes V_1 and V_2 satisfy the relation R_1, where the order of V_1 and V_2 is not completely interchangeable, it can be changed if and only if the edge between V_1 and V_2 is bidirectional, and it cannot be changed when it is a one-way edge. Assuming that each node contained in $V(G)$ has a unique identifier(label), its function is to ensure that each node or each edge is independent and unique.

3.2 Functional dependency and conditional constraints

Definition 2. Suppose D is a functional dependency set of R, and D contains several functional dependencies. For all nodes and edges in the node set $V(G)$, the attribute set is $Attrs$,

If there is an attribute set $X \& Y \subseteq Attrs$, then one functional dependency of D is defined as $D \cong X \to Y$.

If $a_1[X] = a_2[X]$, there must be $a_1[Y] = a_2[Y]$ in the set Y.

Conditional constraints are added on the basis of functional dependence to enrich the expression range[23]. The definitions of conditional constraints in four graph data scenarios are given below.

Definition 3. (*hard constraint*, $HC(U,a)$). Given a certain attribute set U in node $V(G)$, its value is a certain value a.

Definition 4. (*equality constraint*, $EC(U)$)). Given a certain attribute set U in node $V(G)$, it is agreed that its values are equal.

Definition 5. (*edge-to-vertex constraint*, $EV(R_1,R_2,a)$). Several vertices with the same relationship R_1, they and the edge with another relationship R_2 all point to the same vertex a.

Definition 6. (*semantics constraint*, SC). There is no fixed form, usually common sense or logic in our daily lives, which can be abstracted into corresponding forms according to specific contexts.

4 Conflict resolution and missing completion in the fusion of domain ontology

The conflict resolution and missing completion in the fusion of domain ontology will be illustrated by the following two examples:

Example 1. In Figure1, the relationship between $P1$ (i.e. *Person*1)and $P2$ (i.e. *Person*2) is a colleague relationship(R_1 : *Work-with*), and the relationship between $P1$ and $P3$ (i.e. *Person*3) is the same for the colleague relationship, $P1$ contains $a_4 = "Ali"$ among the 4 attributes (a_1,a_2,a_3,a_4) .Similarly, for the 4 attributes in $P2$, there is $a_4 = "Tencent"$, and $R_1 = R(P1,P2) = "Colleague"$. Therefore, it can be inferred that there is a data conflict between the two nodes. Through analysis, the following two situations can be obtained: One is that the relationship between $P1$ and $P2$ is correct, and the a_4 attribute in $P1$ and $P2$ is obvious There is a data quality problem; the second is that the a_4 attributes of $P1$ and $P2$ are correct, then the relationship R_1 between $P1$ and $P2$ is wrong. In these two cases, the next step of processing will get two different results, which are not as expected. In these two cases, the next step of processing will result in two different results, which is not consistent with the expected only repair result, so it is necessary to deal with the above two situations to make the result unique. The reason for the above two situations is the lack of a standard library that can be referred to, that is, the lack of preconditions to unify the two situations. The solution provided by the article is to establish a standard library for reference, that is, the dependency model.

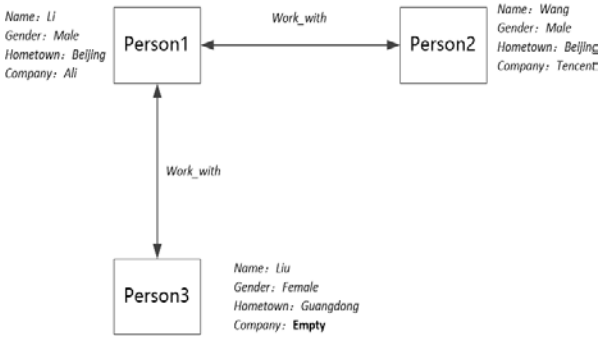

Fig. 1. Directed graph with conflict or missing.

Example 2. As mentioned in Example 1, the relationship between node $P1$ and $P3$ is also R_1 (i.e. colleague relationship) in Figure 1. From node $P3$, it can be intuitively seen that the value of attribute a_4 is empty; obviously, there is a data quality problem with missing data in node $P3$. However, the statement that "the node has a data missing problem intuitively through the naked eye" lacks a convincing basis. Therefore, we compare Example 1, we know that $P1$ and $P3$ are colleague relationships, the attribute a_4 =*Ali* of $P1$, and the attribute a_4 of $P3$ is empty. Obviously, "R_1 =colleague relationship" cannot exist that the company attribute "*company*" of one person is empty, so that $P3$ has the problem of missing data. It can be further deduced that the missing attribute a_4 value of $P3$ is also "Ali". In this way, a method to determine the missing data problem and the corresponding repair method are obtained. Similarly, it is necessary to rely on the model to determine the preconditions (that is, the attribute a_4 do exist in $P1$) is correct.

As a result, the conflict resolution and missing completion will be proceeded by three steps: First, refine the directed graph in which the fusion ontology is stored. And then do the conflict resolution. Finally, do the missing completion.

4.1 Directed Graph Refinement

The structure of the entire graph can be simply classified into two parts: node and edge. Each node has a unique identifier. The inside of the node contains attributes that describe the node, and the edge stores the relationship between the node and the node. In section 3.1, it is mentioned that the internal attributes of the node are not easy to handle. In order to solve this problem, it is necessary to perform some simple pre-processing on the graph data. Take an example to separate the internal attributes of the original node It becomes an independent node and edge, and only retains the identifier of the original node. In order to facilitate the expression and description, the label is used instead. The refinement process algorithm is given as shown in Algorithm 1. The separated node is the attribute value of the original attribute, The edge is the attribute name of the original attribute.

Algorithm 1. Directed Graph refinement.

Input: Directed graph $G_1(V_1, E_1)$, relational set R_1

Output: Directed graph $G_2(V_3, E_2)$, relational set R_2

1. **for** G_2, V_1, E_1 **do**
2. Initialize these three sets to empty sets.
3. **end for**
4. Make $T_1 = V_1$ and $T_2 = E_1$;
5. **for** all $t_i \in T_1$ **do**
6. $V_1 \leftarrow Attrs(V)$.
7. **end for**
8. $V_3 = V_1 \cup V_2$;
9. **for** all $t_j \in T_2$ **do**
10. $R_2 \leftarrow key[t_j]$.
11. **end for**
12. $R_3 = R_2 \cup R_1$.
13. Add the newly added edge to E_1 as E_2, obtain G_2.

4.2 Functional dependency confidence

The domain ontology data is transformed by the storage relation, and the data is converted into a relational table clustered by type according to the storage scheme, and the original graph data structure is converted into data based on relational storage (i.e. relational data).

The concept of data confidence in AFD (Approximate Functional Dependence) can be used to detect the reliability of functional dependence[24]. If person and work-with are related through foreign keys, and the $work-with$ in the header of the table is a functional dependency. The functional dependence obtained in this way is not necessarily correct, so we adopt the concept of data confidence in the classic AFD to measure whether a functional dependence is credible. After calculating the confidence of a certain functional dependency, determine whether to adopt the functional dependency by specifying the validity range of the functional dependency.

Definition 7. (Confidence). The confidence of the functional dependence φ is denoted as $con(\varphi)$. Let Φ be the set of functional dependence, where $\varphi_1 = (work - with)$ and its form is denoted as $\varphi(start \to end)$. Then the standard form of confidence of the functional dependence $con(\varphi)$ is as follows[25].

$$con(\varphi) = \frac{\sum_{x \in \prod_X (I)} \max(c_{XY}(x, y) : y \in \prod_Y I)}{|I|}$$

Definition 8. (Confidence domain). The range of establishment of all functional dependencies is defined as the confidence domain. The value range of the confidence degree is $\frac{X}{I} \le con(\varphi) \le 1$, and we take the value of the confidence domain not less than the midpoint of the value range, denoted as:

$$con(\varphi) \ge \frac{1}{2}(\frac{X}{I} + 1)$$

Definition 9. (Trustable Functional Dependency). The functional dependency that conforms to the confidence domain is called the trustable functional dependence.

4.3 Conflict resolution and missing completion

After refining the graph structure, we can start the next step, that is, to resolve the missing and conflicts in the fusion of domain ontologies.

The abstract process of resolving the conflict problems is as follows:

- 1. Given a directed graph G as a fusion ontology, its relationship set is R, its TFD set is Φ, and all nodes are put into set B,
- 2. Refining the original directed graph to obtain a new directed graph G_1 and a new set of relations R_1, in the new directed graph G_1, the node corresponding to the node V in the original directed graph is V_1 (through the unique identification symbol),
- 3. Pick one initial node V_2, traverse all the nodes associated with it and put it into the set C,

• 4. Take a node V_3 in the set C and delete V_3 from the set C. Corresponding to V_4 in G_2, for the relationship between all the associated edges of V_1 and V_4, compare each function dependency of the traversal function dependency set, and compare all edges that do not meet the conditional dependencies are reset, so that the relationship that points to the same node corresponds to another relationship that also points to the same node, and update the graph G_1,

• 5. Deal with the hard constraint $HC(U,a)$, reset all the edges in the set U so that they all point to the same node a, and update the graph G_1,

• 6. Process the equivalence constraint $EC(U)$ reset all the edges in the set U so that they all point to the same node, and update the graph G_1,

• 7. Repeat steps 4,5,6 until $C=\varnothing$,

• 8. Repeat steps 3,4,5,6 until $B=\varnothing$,

• 9. Return the directed graph G_1 after dealing with the data conflict problem.

The data conflict repair algorithm is given as shown in Algorithm 2.

Algorithm 2. Conflict Resolution

Input: Directed graph $G_1(V_1,E_1)$ before refinement and directed graph G_2 after refinement

Output: Directed graph G_3

1. **for** G_3, V_1 **do**
2. Initialize G_3 to empty, and initialize V_1 to empty set B.
3. **end for**
4. **for** all $b_i \in B$ and all b_i^j related to b_i **do**
5. $C \leftarrow b_i \cup b_i^j$.
6. **end for**
7. **for** all $c_x \in C$ **do**
8. $HC(U,a)$, $EC(U)$, $EV(b_y,b_z,a)$.
9. **end for**
10. Obtain G_3

Next, fix the problem of missing data. After the graph structure is refined, it can be found that the $company-is$ edge of the $P3$ node points to an empty node. In general, the node can be empty, but in this case, it can be noticed that the $P1$ and $P3$ nodes are also in a $work-with$ relationship, so the empty node is not valid and should be a data missing problem. Similarly, the above hard constraints and conclusion are used here, that is, the $company-is$ of $P3$ node edge should point to Ali node.

Give the process of repairing the missing problem:

• 1. For the directed graph G_1 that has repaired the data conflict problem, traverse all the nodes, find the empty nodes among them, and put them into the set E,

• 2. Take any node V_5 from E and traverse all the nodes associated with V_5. If there is only one node associated with V_5, set it as V_6, and set the relationship between V_5 and V_6 as R_1; traverse all the nodes associated with V_6, select the same nodes with R_1 relationship, compare all the functional dependencies of the dependency set one by one to find out the matching functional dependencies, if there are matching functional dependencies, fill in the

missing part according to the functional dependencies, if not, continue processing hard constraints and equivalent constraints until the conditions are found. If no repair method is found after all the conditions are traversed, the empty node will be marked; if there are multiple nodes associated with V_5, the empty node will be marked directly; delete node V_5 and update directed graph G_1,

- 3. Repeat step 2 until the set E is empty,
- 4. Obtain the directed graph G_4.

The data missing repair algorithm is shown in Algorithm 3.

Algorithm 3. Missing Completion

Input: Directed graph G_3

Output: Directed graph G_4

1. Initialize G_4 to empty, traverse all nodes in G_3 and put all empty nodes into set E.
2. **for** all $e_i \in E$ **do**
3. Detect all empty nodes.
4. If a certain functional dependency (or conditional constraint) is met, the node
5. is inferred based on the functional dependency (or conditional constraint).
6. **end for**
7. Obtain nodes set F and edges set H
8. $G_4 = (F, H)$

Combining the repaired results of the three data quality problems of data conflict, data redundancy and data missing, the article gives the repaired graph data.

From the restored graph data, it can be seen that nodes $P1$, $P2$ and $P3$ have three edges with the same relationship pointing to the same node, which means that these relationships of these nodes are related, and this point is consistent with the conditions specified in the functional dependency set and conditional constraints, and furthermore, it means that the graph data is clean graph data.

5 Summary

In the process of fusion of multi-domain ontology, the problems of conflict and missing are inevitable. In this paper, a method is proposed, which described the domain ontology with a directed graph, thereby transforming the problem of conflict resolution and missing completion between domain ontology into the relationship between directed graphs.

The main contributions are as follows:

Designed the processing flow of conflict resolution and missing completion in the process of multi-domain ontology fusion;

Abstract the multi-domain ontology fusion process into the relationship between directed graphs, and then use the method of dealing with directed graphs to deal with the conflicts and deficiencies that may occur in the process.

The future work mainly includes two aspects: on the one hand, mining more special constraints with the characteristics of directed graph; On the other hand, it is mainly to consider the data quality of other types of non-relational data (such as column-based data, key-value pair data, etc.) can use the same (or similar) method. Among them, other kinds of special constraints characteristics can be considered from the connectivity of directed graph, and weights can also be considered to describe the relationship represented by the edges in the directed graph. Or distinguish the newly generated relationship after the refinement operation to simplify the operation of quality problems. For other types of non-relational

data, take key-value pairs as an example, we can consider the connection between key-value pairs and directed graph. Analogy to the method of operating directed graph, one way of thinking is to consider whether directed graph and key-value pair data can be converted to each other (or to consider the corresponding relationship), and the other is to expand the dependency theory to adapt to new scenarios.

References

1. Qin Yi. Research on CPS Testing Technology Facing Environmental Uncertainty[D]. Nanjing University, 2019.
2. Xie JY, Yang J, Chen YG, Wang HX, Yu PS. A sampling-based approach to information recovery. In: Proc. of the ICDE. Iscataway, NJ: IEEE Computer Society, 2008. 476-485. [doi: 10.1109/ICDE.2008.4497456].
3. Mockus A. Engineering big data solutions. In: Proc. of the Future of Software Engineering. ACM Press, 2014. 85?99.
4. Cai L, Zhu Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal 2015; 1-10.
5. Jin CQ, Liu HP, Zhou AY. Functional dependency and conditional constraint based data repair. Ruan Jian Xue Bao/Journal of Software, 2016,27(7):1671-1684 (in Chinese). \url{http://www.jos.org.cn/1000-9825/5037.htm}.
6. Arnaud Castelltort and Anne Laurent. Exploiting NoSQL Graph Databases and in Memory Architectures for Extracting Graph Structural Data Summaries[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2017, 25(1) : 29.
7. Péter Lehotay-Kéry and Attila Kiss. Process, Analyze and Visualize Telecommunication Network Configuration Data in Graph Database[J]. Vietnam Journal of Computer Science, 2020, 07(01) : 12.
8. Chen Ze et al. Research on Automatic Vulnerability Mining Model Based on Knowledge Graph[J]. International Journal on Artificial Intelligence Tools, 2020, 29(07n08).
9. Li JZ, Wang HZ, Gao H. State-of-the-Art of research on big data usability. Ruan Jian Xue Bao/Journal of Software, 2016, 27(7):1605-1625 (in Chinese). \url{http://www.jos.org.cn/1000-9825/5038.htm}.
10. Hamza Turabieh, Amer Abu Salem, Noor Abu-El-Rub. Dynamic L-RNN recovery of missing data in IoMT applications[J]. Future Generation Computer Systems,2018,89.
11. M. TALHA, A. ABOU EL KALAM,N. ELMARZOUQI. Big Data: Trade-off between Data Quality and Data Security[J]. Procedia Computer Science,2019,151.
12. Danilo Ardagna, Cinzia Cappiello,Walter Samá, Monica Vitali. Context-aware data quality assessment for big data[J]. Future Generation Computer Systems,2018.
13. Maryam Ghasemaghaei, Goran Calic. Can big data improve firm decision quality? The role of data quality and data diagnosticity[J]. Decision Support Systems,2019,120.
14. Fan WF, Geerts F. Capturing missing tuples and missing values. In: Proc. of the ACM SIGMOD-SIGACT-SIGART Symp. on Principles of Database Systems. New York: ACM Press, 2010. 169-178. [doi: 10.1145/1807085.1807109].
15. Fan WF, Geerts F, Lakshmanan LVS, Xiong M. Discovering conditional functional dependencies. IEEE Trans. on Knowledge and Data Engineering, 2011,23(5):683-698. [doi: 10.1109/TKDE.2010.154].

16. Bohannon P, Fan WF, Geerts F, Jia X. Conditional functional dependencies for data cleaning. In: Proc. of the ICDE. Piscataway, 2007. 746-755. [doi: 10.1109/ICDE.2007.367920].
17. Bravo L, Fan WF, Ma S. Extending dependencies with conditions. In: Proc. of the VLDB. 2007. 243-254.
18. Carlo Combi, Pietro Sala. Mining approximate interval-based temporal dependencies[J].Acta Informatica, 2016, Vol.53 (6-8), pp.547-585.
19. Jason Van Hulse and Taghi M. Khoshgoftaar and Amri Napolitano. Evaluating the Impact of Data Quality on Sampling[J]. Journal of Information and Knowledge Management, 2011, 10(3) : 225-245.
20. P. C. SAXENA and D. K. TAYAL. NORMALIZATION IN TYPE-2 FUZZY RELATIONAL DATA MODEL BASED ON FUZZY FUNCTIONAL DEPENDENCY USING FUZZY FUNCTIONS[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2012, 20(1) : 99-138.
21. Tu FF, Zhou MH. Data quality problems in software development activity data. Ruan Jian Xue Bao/Journal of Software, 2019,30(5):1522?1531 (in Chinese). \url{http://www.jos.org.cn/1000-9825/5727.htm}.
22. Gong XQ, Jin CQ, Wang XL, Zhang R, Zhou AY. Data-Intensive science and engineering: Requirements and challenges. Chinese Journal of Computers, 2012,35(8):1-16 (in Chinese with English abstract).
23. Olivier Pivert, Etienne Scholly, Grégory Smits, Virginie Thion,Fuzzy quality-Aware queries to graph databases,Information Sciences,Volume 521,2020,Pages 160-173,ISSN 0020-0255,\url{https://doi.org/10.1016/j.ins.2020.02.035}.
24. Chiang F, Miller RJ. A unified model for data and constraint repair. In: Proc. of the ICDE. Iscataway, NJ: IEEE Computer Society, 2011. [doi: 10.1109/ICDE.2011.5767833].
25. Liu BZ, Wang X, Liu PK, Li SZ, Zhang XW, Yang YJ. KGDB: Knowledge graph database system with unified model and query language. Ruan Jian Xue Bao/Journal of Software, 2021,32(3):781?804 (in Chinese). http://www.jos.org.cn/1000-9825/6181.htm.
26. Zhong Ping, Li zhanhuai, Chen Qun. Functional dependency detection method in relational data [J]. Journal of computer science, 2017,40 (01): 207-222.