# Multiple Data Quality Evaluation and Data Cleaning on Imprecise Temporal Data

Xiaoou Ding(✉)

Harbin Institute of Technology, Harbin 150001, Heilongjiang, China
dingxiaoou@stu.hit.edu.cn

**Abstract.** With data currency issues draw the attentions of both researchers and engineers, temporal data, which describes real world events with time tags in database, is playing a key role in data warehouse, data mining, and etc. At the same time, 4V features of big data give rise to the difficulties in comprehensive data quality management and data cleaning. On one hand, entity resolution methods are faced with challenges when dealing with temporal data. On another hand, multiple problems existing in data records are hard to be captured and repaired. Motivated by this, we address data quality evaluation and data cleaning issues in imprecise temporal data. This project aims to solve three key problems in temporal data quality improvement and cleaning: (1) Determining currency on imprecise temporal data, (2) Entity resolution on temporal data with incomplete timestamps, and (3) Data quality improvement on consistency and completeness with data currency. The purpose of this paper is to address the problem definitions and discuss the procedure framework and the solutions of improving the effectiveness of temporal data cleaning with multiple errors.

**Keywords:** Temporal data · Data currency · Multiple data cleaning
Data quality

## 1 Introduction

Data quality evaluation and data cleaning plays a key role in the whole life circle of big data management today [23]. With the increasing demand in efficiency and currency, the *temporal* feature in data has been recognized as an important issue in data science. However, the quality problems in temporal data are often quite serious and trouble data transaction steps (e.g., acquisition, copy, querying). On one hand, various information systems store data with different formats, which makes entity resolution process necessary. The attribute values are possible to change with the evolution of time. It adds to the difficulties in entity resolution. In practice, certain attribute of records referring to the same entity may change over time. All of these records may be valid and proper for describing a certain entity only at a particular time period. For example, DBLP


C. Woo et al. (Eds.): ER 2018 Workshops, LNCS 11158, pp. 69–75, 2018.
https://doi.org/10.1007/978-3-030-01391-2_14

collects researchers' paper information for many years. The information for the same author about different papers may be different since the author's affiliation, partners or even names may change over time. On another hand, currency, consistency and completeness problems are always costly in multi-source data integration. These problems result in the low reliability of data, and they also add to the confusion in data applications.

Researchers have gone a long way in data quality and data cleaning, particularly in accuracy, consistency, completeness and record de-duplication [3,12,14]. It has been found that currency issues in temporal data also seriously impact data repairing. In addition, these problems are possible to affect each other during repairing, rather than completely isolated [5,14]. However, existing data repairing methods fails to pay attention to the temporal feature in data, without which the performance of them is challenged to break through bottleneck. Motivated by this, we propose multiple data cleaning in temporal data in this project.

## 2 Problem

We address two main issues in this project, as summarized below.

**Problem 1:** We first study the entity resolution method on imprecise temporal data. Due to the fact that reliable timestamps are often not available and complete, or even absent in practice, it is necessary to improve the entity resolution method to perform well on imprecise temporal data. We propose the problem definition in Definition 1.

**Definition 1.** *$\mathcal{R} = (A_1, ...A_m)$ is a relation schema and $I_R = \{r_1, ..., r_n\}$ is a set of instances of $\mathcal{R}$ which contains n records. $I_R$ has no precise timestamps.*

*The **Entity Resolution problem on imprecise temporal data** is to cluster the records into clusters, such that records in the same clusters refer to the same entity over time and records in different clusters refer to different entities.*

**Problem 2:** We study the detecting and repairing approach on incomplete and inconsistent data with incomplete timestamps, to achieve multiple data quality improvement on completeness, consistency and currency. The problem definition is presented in Definition 2.

**Definition 2.** ***Data cleaning on consistency, completeness and currency**. Given a low-quality data $\mathcal{D}$, data quality rules including a set $\Phi$ of* CC*s and a set $\Sigma$ of* CFD*s, and a confidence $\sigma$ for each attributes. Data quality improvement on $\mathcal{D}$ with completeness, consistency and currency is to detect the dirty data in $\mathcal{D}$ and repair it into a clean one, denoted by $\mathcal{D}_r$, where*

*(a) $\forall r(r \in \mathcal{D}_r)$ has a reliable currency order value satisfying the set $\Phi$ of* CC*s, denoted by $(\mathcal{D}_r, \mathcal{D}) \models \Phi$.*
*(b) $\mathcal{D}_r$ is consistent referring to the set $\Sigma$ of* CFD*s, i.e., $(\mathcal{D}_r, \mathcal{D}) \models \Sigma$.*
*(c) The missing values in $\mathcal{D}$ are repaired with the clean ones whose confidence $> \sigma$ into $\mathcal{D}_r$.*
*(d) The repair cost $cost(\mathcal{D}_r, \mathcal{D})$ is as small as possible.*

## 3 Proposed Framework and Solution

**Solution of Problem 1**: The method overview is shown in Fig. 1, which has two main parts: similarity comparison and clustering, respectively. We first define temporal attributes' unstableness, which depicts the evolving probability of attribute values of entities. Accordingly, a dynamic weight scheme is designed to match these records more precisely than the traditional fixed weight scheme.

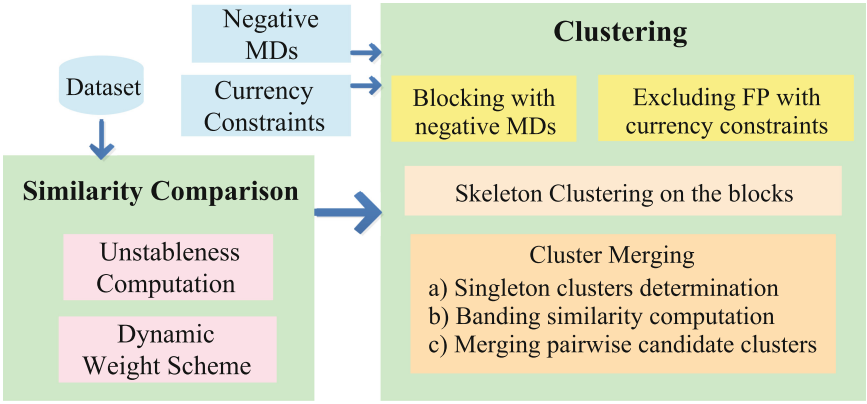


**Fig. 1.** System architecture for Problem 1

Then, we combine the data currency constraint and matching dependency rules effectively and propose a temporal clustering algorithm that takes data currency into consideration to guarantee result quality.

After that, we develop a rule-conducted uniform framework for resolving temporal records with integration of data quality rules, and design several efficient algorithms to track the problems in the framework. Specifically, we propose algorithms to (1) block records into disjoint blocking and exclude false positives in each block, (2) compute unstableness of attributes and dynamic weight scheme, (3) generate a cluster of each block, and (4) determine initial clusters in the clustering results.

**Solution of Problem 2**: Since that completeness and consistency are metrics focusing on measuring the quality with features in values, while currency describes the temporal order or the volatility of records in the whole data set. We process consistency and completeness repairing along the currency order, respectively. We outline the 3C data cleaning method in Fig. 2. It includes four main steps.

**Step 1**: We first construct currency graphs for records with the given CCs, and make conflict resolution in the currency graphs. If conflicts exist, the conflicted CCs and the involved records will be returned. They are supposed to be fixed by domain experts or revised from business process.

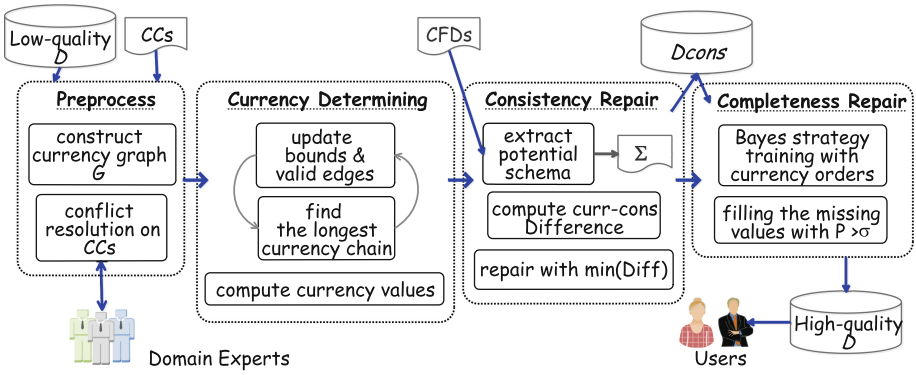


**Fig. 2.** Framework for Problem 2

**Step 2**: We then determine the currency order of all the records extracted from CCs. We update the longest currency order chain in each currency graph iteratively, and compute currency values to each record. This currency order is obtained as a direct and unambiguous metric among records on currency.

**Step 3**: After that, we repair consistency issues with the global currency orders. We input consistency constraints (CFDs for short) first, and then we define a distance metric *Diffcc* to measure the distance between dirty data and clean ones, combining consistency difference with currency orders. That is, we attempt to repair the dirty data with proper values which have the closest current time point.

**Step 4**: We repair incomplete values with Bayesian strategy because of its obvious advantages in training both discrete and continuous attributes in relational database. We treat currency orders as a weighted feature and train the complete records to fill in the missing values if the filling probability no less than a confidence measure. Up till now, we achieve high-quality data on 3C.

## 4 Related Work

Study on data quality is extensive for decades in many fields. With the demand for high quality data, many metrics beside accuracy are necessary for quality improvement [27].

**Data Quality Evaluation.** [3] provides a systematic introduction of data quality methodologies, in which completeness, accuracy, consistency and currency are four important dimensions. As for completeness issues, many kinds of algorithms have been proposed to fill the missing values [10], like statistic-based, probability-based, and learning oriented and etc. As for consistency, different semantic constraints such as FD, CFD, and CIND, have been defined to guide data cleaning under specific circumstance [4,12], where conditional functional

dependency (CFD) is a general and effective consistency constraints for querying and inconsistency detection in database [9,13]. Also, consistency constraints as well as data quality rules discovery problem is proposed and studied in works like [7,17]. Currency describes to which extent a data set is up-to-date [3]. When various data sources are integrated, timestamps are always neither complete nor uniform. It promotes the study on currency determination without available timestamps. [15] is the first to propose a constraint-based model for data currency reasoning. And several fundamental theoretical problems are discussed in both [12,15].

Furthermore, as the dimensions are not independent issues in data integration [12], data cleaning approaches have been developed with integrating several data quality dimensions. [14] reports advanced study on critical dimensions and provides a logical framework for a uniform treatment of the issues. [9] propose a framework for quality improvement on both consistency and accuracy. [5] discusses time-related measures with accuracy and completeness, and proposes functions of computing their mutual relationships.

**Entity Resolution.** (ER) also known as record linkage, duplicate detection, and duplicate identification, aims to take a set of records as input and find out the records referring to the same real-world entities [2,18,26,28]. Researchers have developed multiple similarity metrics for matching duplicate records, including character-based [1,24], token-based [8] and numeric similarity metrics [19]. [11] summarizes several common categories of ER, including probabilistic matching [25], supervised learning [22], rule-based approaches [19,20], etc.

With temporal data accumulated rapidly in variety information systems, some methods coped with lexical heterogeneity in ER problem may fail to perform well in resolving temporal records directly due to existence of evolving heterogeneity. [21] studies linking temporal records and propose time decay to capture the evolution of entity value over time. [16,28] study efficient rule evolution techniques for clustering issues in ER. [6] proposes a fast algorithm to match temporal records.

While advanced techniques in temporal ER problems achieve high efficiency and accuracy, few methods can be well applied to record matching on datasets without timestamps. The reason is that existing methods mostly depend on definite timestamps and under such circumstance, we can only reason a relative currency order with currency constraints. It motivates us to propose a rule-based ER method to address the entity value evolution effectively on imprecise temporal data.

## 5 Conclusion and Future Work

Now, we are in the midst of cleaning temporal data with multiple errors. We have already propose an entity resolution method in imprecise temporal data. We propose attribute unstableness to capture the entity evolution over time, and apply dynamic weight schema for improving pairwise similarity computation. We apply rules to determine the currency order of records from target attributes, and

propose a novel clustering algorithm along with a pruning method to improve the quality of ER.

We have constructed currency order graph with currency constraints, and determined the currency of each record, accordingly. Based on the currency orders deduced from currency constraints, we will construct the complete method of repairing low-quality data with incomplete and inconsistent values, which lacks for available timestamps. Future works includes (1) seeking for more rules to model the evolving trend of the temporal data more accurate and learning efficient methods to find the rules, and (2) design the balance strategy of cleaning cost and effectiveness of the multiple data cleaning.

## References

1. UNIMATCH: a record linkage system: users manual. In: Bureau of the Census, Washington DC (1976)
2. Ananthakrishna, R., Chaudhuri, S., Ganti, V.: Eliminating fuzzy duplicates in data warehouses. In: International Conference on Very Large Data Bases, pp. 586–597 (2002)
3. Batini, C., Cappiello, C., Francalanci, C., Maurino, A.: Methodologies for data quality assessment and improvement. ACM Comput. Surv. **41**(3), 16 (2009)
4. Bertiequille, L., Sarma, A.D., Dong, Marian, A., Srivastava, D.: Sailing the information ocean with awareness of currents: discovery and application of source dependence. Computer. Science **26**(8), 1881–3 (2009)
5. Cappiello, C., Francalanci, C., Pernici, B.: Time related factors of data accuracy, completeness, and currency in multi-channel information systems. In: The Conference on Advanced Information Systems Engineering, pp. 145–153 (2008)
6. Chiang, Y.H., Doan, A.H., Naughton, J.F.: Tracking entities in the dynamic world: a fast algorithm for matching temporal records. Proc. VLDB Endow. **7**, 469–480 (2014)
7. Chu, X., Ilyas, I.F., Papotti, P., Ye, Y.: Ruleminer: data quality rules discovery. In: IEEE International Conference on Data Engineering, pp. 1222–1225 (2014)
8. Cohen, W.W.: Integration of heterogeneous databases without common domains using queries based on textual similarity. In: ACM SIGMOD International Conference on Management of Data, pp. 201–212 (1998)
9. Cong, G., Fan, W., Geerts, F., Jia, X., Ma, S.: Improving data quality: consistency and accuracy. In: International Conference on Very Large Data Bases, pp. 315–326 (2007)
10. Deng, T., Fan, W., Geerts, F.: Capturing missing tuples and missing values. In: Twenty-Ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2010, Indianapolis, Indiana, USA, 6–11 June 2010, pp. 169–178 (2010)
11. Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S.: Duplicate record detection: a survey. IEEE Trans. Knowl. Data Eng. **19**(1), 1–16 (2007)
12. Fan, W., Geerts, F.: Foundations of Data Quality Management (2012)
13. Fan, W., Geerts, F., Jia, X.: Conditional dependencies: a principled approach to improving data quality. In: Sexton, A.P. (ed.) BNCOD 2009. LNCS, vol. 5588, pp. 8–20. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02843-4_4

14. Fan, W., Geerts, F., Ma, S., Tang, N., Yu, W.: Data quality problems beyond consistency and deduplication. In: Tannen, V., Wong, L., Libkin, L., Fan, W., Tan, W.-C., Fourman, M. (eds.) In Search of Elegance in the Theory and Practice of Computation. LNCS, vol. 8000, pp. 237–249. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41660-6_12
15. Fan, W., Geerts, F., Wijsen, J.: Determining the currency of data. ACM Trans. Database Syst. **37**(4), 71–82 (2012)
16. Fan, W., Jia, X., Li, J., Ma, S.: Reasoning about record matching rules. Proc. VLDB Endow. **2**(1), 407–418 (2009)
17. Fei, C., Miller, R.J.: A unified model for data and constraint repair. In: IEEE International Conference on Data Engineering, pp. 446–457 (2011)
18. Fellegi, I.P., Sunter, A.B.: A theory for record linkage. J. Am. Stat. Assoc. **64**(328), 1183–1210 (1969)
19. Koudas, N., Marathe, A., Srivastava, D.: Flexible string matching against large databases in practice. In: Thirtieth International Conference on Very Large Data Bases, pp. 1078–1086 (2004)
20. Li, L., Li, J., Gao, H.: Rule-based method for entity resolution. IEEE Trans. Knowl. Data Eng. **27**(1), 250–263 (2015)
21. Pei, L.I., Dong, X.L., Maurino, A., Srivastava, D.: Linking temporal records. PVLDB **4**(11), 956–967 (2011)
22. Richman, J., Richman, J.: Learning to match and cluster large high-dimensional data sets for data integration. In: Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 475–480 (2002)
23. Sidi, F., Panahy, P.H.S., Affendey, L.S., Jabar, M.A., Ibrahim, H., Mustapha, A.: Data quality: a survey of data quality dimensions. In: International Conference on Information Retrieval and Knowledge Management, pp. 300–304 (2012)
24. Ullmann, J.R.: A binary n-gram technique for automatic correction of substitution, deletion, insertion and reversal errors in words. Comput. J. **20**(2), 141–147 (1977)
25. Verykios, V.S., Moustakides, G.V., Elfeky, M.G.: A bayesian decision model for cost optimal record matching. VLDB J. **12**(1), 28–40 (2003)
26. Verykios, V.S., Elmagarmid, A.K., Houstis, E.N.: Automating the approximate record-matching process. Inf. Sci. **126**(1–4), 83–98 (2002)
27. Wang, R.Y., Strong, D.M.: Beyond accuracy: what data quality means to data consumers. J. Manag. Inf. Syst. **12**(4), 5–33 (1996)
28. Whang, S.E., Garcia-Molina, H.: Entity resolution with evolving rules. Proc. VLDB Endow. **3**(1–2), 1326–1337 (2010)