Contextual Data Cleaning

Morteza Alipour Langouri #1, Zheng Zheng #2, Fei Chiang #3, Lukasz Golab *4, Jaroslaw Szlichta ÷5

# McMaster University, Canada, * University of Waterloo, Canada, ÷ University of Ontario Institute of Technology, Canada
1 alipoum@mcmaster.ca. 2 zhengz13@mcmaster.ca, 3 fchiang@mcmaster.ca,
4 lgolab@uwaterloo.ca, 5 jaroslaw.szlichta@uoit.ca

***Abstract*—In this paper, we motivate the need to include context in data cleaning in order to account for the subjective nature of data quality. Based on our recent work on incorporating ontologies into Functional Dependencies, we argue that ontologies are a rich source of context, and an effective tool for modeling domain concepts and relationships for data cleaning. Using real datasets, we present examples showing how ontologies can improve data cleaning workflows, and we outline open problems and directions for future work.**

I. Introduction

Data cleaning has been an active area of data management research and practice, motivated by the observation that real datasets are rarely error-free. Errors may be caused by improper data collection and data entry, unknown or undocumented data semantics, or malfunctioning or misconfigured devices.

Data cleaning workflows typically consist of two steps: identifying values that are potentially incorrect, and suggesting possible modifications or *repairs* for them. Declarative or qualitative data cleaning has been the most popular approach in the literature. Declarative data cleaning techniques use integrity constraints to define the specific attribute relationships that the data should satisfy. These integrity constraints (also referred to as data dependencies; e.g., Functional Dependencies (FDs)) serve as a benchmark to repair data values such that the data and dependencies are consistent [1], [2], [3], [4], [5], [6]. In these settings, the definition of cleanliness is clear: if the data satisfies the dependencies, then it is considered clean, otherwise it is considered dirty.

Recent efforts to expand data cleaning semantics include holistic based cleaning that are amenable to a variety of data quality rules [7], [8], [9], and statistical cleaning, which proposes updates to the data according to expected statistical distributions [10], [11], [12]. All these techniques aim to make corrections to the data by optimizing a predefined cost function that selects the best repair(s) to the data from a space of candidate repairs.

However, data quality is known to be subjective and highly contextual based on individual preferences, domain specific definitions, or regional conventions. For example, medical drug names can vary based on whether brand name versus generic name is used, and the geographic region where the drug is prescribed. The drug `Paracetamol`, commonly known as `Acetaminophen` in North America, is a painkiller marketed under different brand names, such as `Tylenol` in North America and `Panadol` in the United Kingdom. Domain expertise is often required to supplement the mere description of the data; in this example, to reconcile that these differing syntactic references refer to the same entity. This contextual information provides rules, language and conventions related to the domain (application or industry) that is essential to correctly interpret the data.

TABLE I: Sample clinical trials data

id	CC	CTRY	SYMP	DIAG	MED
t_1	US	United States	joint pain	osteoarthritis	ibuprofen
t_2	IN	India	joint pain	osteoarthritis	NSAID
t_3	CA	Canada	joint pain	osteoarthritis	naproxen
t_4	IN	Bharat	nausea	migrane	analgesic
t_5	US	America	nausea	migrane	tylenol
t_6	US	USA	nausea	migrane	acetaminophen
t_7	IN	India	chest pain	hypertension	morphine

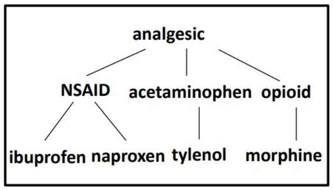

Fig. 1: A medical ontology

Consider the following example of using FDs for error identification where context is needed to resolve ambiguous references. Table I shows a real data sample of clinical records from the Linked Clinical Trials database from *http://www.linkedmdb.com*, containing patient country codes (CC), country (CTRY), symptoms (SYMP), diagnosis (DIAG), and the prescribed medication (MED). Consider two FDs: F_1: [CC] $\rightarrow$ [CTRY] and F_2: [SYMP, DIAG] $\rightarrow$ [MED]. The tuples (t_1, t_5, t_6) violate F_1 as 'United States', 'America', and 'USA' are *not syntactically equivalent* (the same is true for (t_2, t_4, t_7)). However, 'United States' is synonymous with 'America' and 'USA', and (t_1, t_5, t_6) all refer to the same country. Similarly, 'Bharat' in t_4 is synonymous with 'India' as it is the country's original Sanskrit name. For F_2, (t_1, t_2, t_3) and (t_4, t_5, t_6) do not satisfy the dependency as the right-hand-side values all refer to different medications. However, with domain knowledge from a medical ontology shown in Figure 1, we see that the values participate in an inheritance (is-a) relationship. Both 'ibuprofen' and 'naproxen' are non-steroidal anti-inflammatory drugs (NSAID), and 'tylenol' is an 'acetaminophen' drug, which in turn is an 'analgesic'.

In the above example, having context is essential towards

2473-3490/18/$31.00 ©2018 IEEE
DOI 10.1109/ICDEW.2018.00010

correct entity reconciliation. Existing dependencies such as FDs fail to recognize that different syntactic references to an entity may be semantically equivalent. This consequently leads to an increased number of false positive errors. In F_1, we initially identify three "errors" with different references to the 'United States', and similarly three "erroneous" tuples with different references to India. If contextual information about synonyms is available, then we recognize that these six tuples ($t_1, t_2, t_4 - t_7$) are in fact all correct. This reduces the computational effort of data cleaning, and the manual burden for users to identify falsely categorized data errors.

Unfortunately, *context* (domain knowledge) is often only available via user-expertise where manual data cleaning is performed. If we rely on human effort for large datasets, false positive errors will inevitably still remain. Automated solutions are needed to adequately model and capture context to improve the accuracy of data cleaning results.

We argue that *ontologies* are a rich source of context, and an effective tool for modeling domain concepts and relationships for data cleaning. Ontologies are under-utilized in existing dependency-based cleaning techniques due to a lack of formalism to integrate ontologies with data quality dependencies. We start with a review of related work in Section II. Next, we present examples from real datasets demonstrating the utility of ontologies in data cleaning to reduce the number of false positive errors, and to improve repair accuracy (Section III). In Section IV, we outline a set of challenges to be addressed to achieve a contextual data cleaning pipeline. We also summarize our initial work in this area by introducing a novel class of dependencies called *Ontology Functional Dependencies (OFDs)* and the effective OFD discovery algorithm [13]. This lays the groundwork for integrating context with data dependencies. We conclude in Section V.

II. Related Work

Previous work on injecting context into data cleaning has focused on human-in-the-loop solutions. One solution is to leverage *master data* curated by domain experts. Prior work on data cleaning with master data considers tuple-level pattern matching [14]. For example, an address master dataset may indicate that if zip code = 10001 then city = New York; this can be used to determine the correct city attribute for all rows where zip code = 10001. Using richer contextual information such as that found in ontologies was not considered.

Several techniques involve humans during the data cleaning process. These include crowd sourcing to bring human insight during record matching [15], and for identifying persistent errors that may remain after cleaning [16]. User-guided selection of repairs include examples such as GDR [17], KATARA [9] and UGuide [18]. Here, one important technical problem is to determine which values should be verified by experts assuming that a bounded number of such verifications can be performed. Returning to our medical example, human experts might be aware of the synonym relationships and might be able to resolve some of the inconsistencies. We advocate for a deeper integration of context into the data cleaning framework so that such "inconsistencies" are not flagged in the first place, minimizing the need for post-cleaning, crowd sourced error detection techniques [16].

To the best of our knowledge, the only prior work on using ontologies for data cleaning is KATARA. However, KATARA focuses on simple patterns in ontologies such as "France" hasCapital "Paris", and does not integrate ontologies into the definitions of integrity constraints. We will describe our recent work on Ontology Functional Dependencies [13] in Section IV.

III. Motivating Examples

Ontologies are an effective tool to capture domain knowledge where entities, their properties, and relationships are modeled via a given taxonomy. An ontology S is an explicit specification of a domain that includes concepts, entities, properties, and relationships among them. These constructs are often defined and applicable only under a given context (or interpretation), called a *sense*. The meaning of these constructs for a given S can be modeled according to different contexts, leading to different ontological interpretations. For example, the value '`jaguar`' can be interpreted under two senses: (1) as an animal, and (2) as a vehicle. As an animal, '`jaguar`' is synonymous with '`panthera onca`', but not with '`jaguar land rover`' which is an automotive company. By having the appropriate context, data cleaning operations can be correctly interpreted according to given ontological concepts and relationships. Returning to our medical example, having context of the geographic region where a drug is used, such as in North America, can disambiguate `Acetaminophen` and `Tylenol`, which are synonyms and reference the same drug.

To quantify the benefits of including ontology relationships in data cleaning, we profiled three datasets: clinical trials data from the Linked Clinical Trials database (1M records), pollution data (460K records) from the Canadian Pollutant Release Inventory, and census-income data (300K records) from the UCI Machine Learning Repository [19]. Using medical [20] and country code [21] ontologies, our objective was to quantify the benefit of including *synonym* and *is-a* (inheritance) relationships defined in the ontology during the error detection process using FDs. We defined a set of FDs of the form $X \to Y$ over each dataset, where X is a set of attributes in the schema and Y is a single attribute. We then computed the percentage of tuples that participated in either a synonym (SYN) or inheritance (ISA) relationship in the Y attribute, under a given context. Under traditional FD-based data cleaning approaches, which consider syntactic equality, these identified tuples would be falsely considered erroneous.

Table II shows that a significant percentage of tuples from each dataset participate in synonym or inheritance relationships. In the clinical data, for FD F_1: [disease] → [medicine], 37.6% and 44.1% of tuples participate in a synonym relationship under the 'US' and 'EU' context. We sampled the data and observed examples such as 'asthma' is treated by medications {'Advicor', 'Advair', 'Seretide'}, which are

TABLE II: False positives identified via ontologies.

Dataset	**Rel**	Y	**Context**	**%**
Clinical	SYN	CountryCode	ISO	48.2%
			UN	53.9%
	ISA	Disease	Med	20.9%
	SYN	Medicine	US	37.6%
			EU	44.1%
Census	SYN	NativeCountry	ISO	66.2%
			UN	69.7%
Pollution	SYN	Province	ISO	63%
			UN	71.1%

all synonymous in the 'EU' context. However, in the 'US', 'Seretide' is not recognized as a synonym, and is flagged as an error. Other discovered cases include country code names such as {Germany, Deutschland, DE, DEU} which are synonyms under conventions used by the United Nations (UN), but not by the International Organization for Standardization (ISO). Table II reveals that provincial and country name variations account for 48%-71% of false positive errors that can be eliminated by injecting context from ontologies.

Unfortunately, the use of ontologies in data cleaning has been limited, and existing data dependencies cannot capture the semantics modeled in an ontology. As a result, dependency based data cleaning systems have a common flaw: they incorrectly label syntactically different but semantically equivalent values as errors. This leads to an increased number of reported "errors" and a larger search space of candidate data repairs.

IV. Challenges

We now discuss a set of open problems to achieve contextual data cleaning and summarize our work to address some of these problems.

A. *Integrating Ontologies with Data Dependencies*

The first challenge requires defining a new class of *contextual* dependencies that integrate ontological concepts and relationships with data dependencies. In dependency based data cleaning systems, this will provide the necessary context for more accurate error detection and repair. This first involves identifying an existing class of dependencies to extend with context, and selecting meaningful ontological relationships and properties to apply. In our use case study, we explored the *synonym* and *inheritance* ontological relationships, which have shown to be prevalent in real data. While a suite of data dependencies have been proposed in recent years for data cleaning, FDs have remained the most commonly used thus far. The integration of context defines a new class of dependencies that involves the study of their properties, axioms, relationship to existing data dependencies, and efficient algorithms for discovering these contextual dependencies.

Towards this effort, we have defined a new class of dependencies called *Ontology Functional Dependencies (OFDs)* that integrates synonym and inheritance relationships from an ontology with FDs [13]. Intuitively, an OFD $X \rightarrow Y$, paired with a corresponding ontology, asserts that if any set of tuples agree on all the attributes in X, then there must exist some interpretation (context) under which all the Y values are synonyms or participate in an is-a relationship.

Recall that the dataset in Table I violates the FDs [CC] → [CTRY] and [SYMP, DIAG] → [MED]. However, the corresponding OFDs, assuming the ontology from Figure 1, are not violated and therefore no errors are flagged. For example, tuples (t_1,t_5,t_6) no longer violate the first dependency because there exists a common interpretation (country name) under which "United States", "America" and "USA" are synonyms.

Notably, OFDs *cannot* be reduced to traditional FDs, making them interesting and non-trivial to work with. Since values may have multiple senses (e.g., jaguar the animal and jaguar the car), it is not possible to create a normalized table by replacing each value with a unique canonical name.

We identify at least two directions for future work on integrating ontologies with data dependencies:

1) *Studying other types of dependencies that include false positive errors without considering context. These dependencies include Conditional Functional Dependencies [22], Inclusion Dependencies [1], Order Dependencies [23] and Denial Constraints [5].*
2) *Deeper integration of ontologies: other ontological relationships beyond synonyms and inheritance, such as part-of. Extensions to include ontological relationships among the left-hand-side attributes to capture a greater number of true positive errors.*

B. *Discovery of Contextual Data Quality Rules*

After formulating a suite of contextual data quality rules, we can use them to identify erroneous data. However, manual specification of data dependencies is costly because data semantics are often poorly documented and change over time. In systems with hundreds or thousands of tables, with multiple ontologies defined over a table, automated solutions are needed to discover contextual data quality rules.

The possibility of having multiple ontological contexts defined over a data instance poses additional complexity (over existing dependency discovery algorithms) as the data quality rule is valid under a specific context. That is, values participating in a discovered rule are interpreted under a specific context. In our earlier example, we can define a contextual data quality rule ϕ_1: [disease] → [medicine] that holds under the "EU" context, but holds approximately (with some errors) under the "US" context due to differing synonyms between the two geographies.

In our recent work, we proposed an algorithm that discovers OFDs from data [13]. To illustrate the additional complexity due to the possibility of multiple contexts, recall that identifying violations of a traditional FD $X \rightarrow Y$ is simple: it suffices to check all pairs of tuples agreeing on the X attributes and ensure that they agree on Y. However, verifying whether an OFD holds is more involved because we must find a common interpretation of the Y-values for each set of tuples agreeing on X. As a result, existing dependency discovery algorithms that validate dependencies via pairwise tuple comparisons (e.g., FastFD [24]) cannot be easily and efficiently extended to OFDs. Instead, the algorithm we proposed uses an Apriori-like approach, similar to the TANE FD discovery algorithm [25], along with novel optimizations to prune the search space of possible rules.

We identify two directions for future work on discovering contextual rules:

1) *Discovering other types of dependencies augmented with ontological context.*
2) *Ranking the discovered dependencies. This is an important problem because the number of such dependencies may be large and some of them may be specific to the given data instance (i.e., overfitting).*

C. Contextual Data Cleaning

Given a class of contextual data quality rules and efficient discovery algorithms, applying these contextual dependencies for error detection naturally requires context aware data cleaning algorithms. In general, data cleaning algorithms assume that the dataset is mostly clean and attempt to minimize the number of changes. A simple such algorithm may take a frequency based approach. In this approach, infrequent values are updated to frequent ones based on the assumptions that frequent values serve as positive evidence of clean data.

However, we envision a more holistic view of context-aware cleaning that considers repairs to the data, to the contextual dependencies, and the ontology/knowledge base itself (there has been work on holistic data cleaning that considers data modifications as well as FD modifications [4], [2], but ontologies were not considered). Such a model requires defining a comparable notion of an error among the data, the dependency and the ontology, and defining a cost model to quantify the degree of error in each case. Finally, data cleaning techniques are needed that evaluate the space of possible repairs under different ontological contexts, and defining the conditions for an optimal repair.

To summarize, open problems in contextual data cleaning include the following:

1) *Developing algorithms that suggest possible modifications of data that violate a given set of contextual dependencies.*
2) *Developing holistic data cleaning algorithms that simultaneously consider data, dependency and ontology repairs.*

V. Conclusions

In this paper, we motivated the need to study contextual data cleaning, in which ontologies are integrated into data dependencies and cleaning decisions. Compared to existing dependencies and data cleaning workflows, contextual data cleaning models richer relationships in the data and avoids misclassifying consistent data as erroneous. We summarized our recent work on Ontology Functional Dependencies and suggested directions for future research.

We remark that ontologies are not the only way of specifying context, and other techniques for including context in data cleaning should also be studied. Furthermore, while contextual dependencies are unlikely to eliminate the need for human involvement in the data cleaning process, they have the potential to reduce the burden of human verification.

References

[1] P. Bohannon, M. Flaster, W. Fan, and R. Rastogi, "A cost-based model and effective heuristic for repairing constraints by value modification," in *SIGMOD*, 2005, pp. 143–154.

[2] F. Chiang and R. J. Miller, "A unified model for data and constraint repair," in *ICDE*, 2011, pp. 446–457.

[3] F. Chiang and R. Miller, "Active repair of data quality rules," in *IJIQ*, 2011, pp. 174–188.

[4] G. Beskales, I. F. Ilyas, L. Golab, and A. Galiullin, "On the relative trust between inconsistent data and inaccurate constraints," in *ICDE*, 2013, pp. 541–552.

[5] X. Chu, I. F. Ilyas, and P. Papotti, "Holistic data cleaning: Putting violations into context," in *ICDE*, 2013, pp. 458–469.

[6] M. Volkovs, F. Chiang, J. Szlichta, and R. J. Miller, "Continuous data cleaning," in *ICDE*, 2014, pp. 244–255.

[7] M. Dallachiesa, A. Ebaid, A. Eldawy, A. Elmagarmid, I. F. Ilyas, M. Ouzzani, and N. Tang, "NADEEF: A commodity data cleaning system," in *SIGMOD*, 2013, pp. 541–552.

[8] F. Geerts, G. Mecca, P. Papotti, and D. Santoro, "The LLUNATIC data-cleaning framework," *PVLDB*, vol. 6, no. 9, pp. 625–636, 2013.

[9] X. Chu, J. Morcos, I. F. Ilyas, M. Ouzzani, P. Papotti, N. Tang, and Y. Ye, "Katara: A data cleaning system powered by knowledge bases and crowdsourcing," in *SIGMOD*, 2015, pp. 1247–1261.

[10] L. Berti-Equille, T. Dasu, and D. Srivastava, "Discovery of complex glitch patterns: A novel approach to quantitative data cleaning," in *ICDE*, 2011, pp. 733–744.

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

[12] N. Prokoshyna, J. Szlichta, F. Chiang, R. J. Miller, and D. Srivastava, "Combining quantitative and logical data cleaning," *Proc. VLDB Endow.*, vol. 9, no. 4, pp. 300–311, Dec. 2015.

[13] S. Baskaran, A. Keller, F. Chiang, L. Golab, and J. Szlichta, "Efficient discovery of ontology functional dependencies," in *CIKM*, 2017, pp. 1847–1856.

[14] W. Fan, J. Li, S. Ma, N. Tang, and W. Yu, "Towards certain fixes with editing rules and master data," *PVLDB*, vol. 3, no. 1, pp. 173–184, 2010.

[15] J. Wang, T. Kraska, M. J. Franklin, and J. Feng, "Crowder: Crowdsourcing entity resolution," *Proc. VLDB Endow.*, vol. 5, no. 11, pp. 1483–1494, 2012.

[16] Y. Chung, S. Krishnan, and T. Kraska, "A data quality metric (dqm): How to estimate the number of undetected errors in data sets," *Proc. VLDB Endow.*, vol. 10, no. 10, pp. 1094–1105, Jun. 2017.

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

[18] S. Thirumuruganathan, L. Berti-Equille, M. Ouzzani, J.-A. Quiane-Ruiz, and N. Tang, "Uguide user-guided discovery of fd-detectable errors," in *SIGMOD*, 2017, pp. 1385–1397.

[19] "Uci machine learning repository." [Online]. Available: http://archive.ics.uci.edu/ml/index.php

[20] "Institute for safe medication practices." [Online]. Available: http://www.ismp.org/tools/default.aspx

[21] "International organization for standardization, country codes." [Online]. Available: https://www.iso.org/iso-3166-country-codes.html

[22] G. Cong, W. Fan, F. Geerts, X. Jia, and S. Ma, "Improving data quality: Consistency and accuracy," in *VLDB*, 2007, pp. 315–326.

[23] J. Szlichta, P. Godfrey, L. Golab, M. Kargar, and D. Srivastava, "Effective and Complete Discovery of Order Dependencies via Set-based Axiomatization," *PVLDB*, vol. 10, no. 7, pp. 721–732, 2017.

[24] C. Wyss, C. Giannella, and E. L. Robertson, "FastFDs: Heuristic-driven, depth-first alg. for mining FDs from relations," in *DaWaK*, 2001, pp. 101–110.

[25] Y. Huhtala, P. P. J. Kinen, and H. Toivonen, "Efficient discovery of functional and approximate dependencies using partitions," *ICDE*, pp. 392–401, 1998.