Leopoldo Bertossi1⋆⋆, Flavio Rizzolo2⋆⋆⋆, Lei Jiang3

1 Carleton University, Ottawa, Canada. bertossi@scs.carleton.ca

2 Carleton University, Ottawa, Canada. flavio@scs.carleton.ca

3 University of Toronto, Toronto, Canada. [leijiang@cs.toronto.edu](mailto:leijiang@cs.toronto.edu)

**Streszczenie** Uzasadniamy, formalizujemy i badamy takie pojęcia jak ocena jakości danych oraz odpowiadanie na zapytania dotyczące jakości danych jako działania zależne od kontekstu. Konteksty dla oceny i wykorzystania podręcznych źródeł danych są modelowane jako zewnętrzne bazy danych (które mogą być zmaterializowane albo wirtualne) i mapowania wewnątrz zbiorów oraz podręczne źródła danych. W ten sposób, kontekst staje się "uzupełnieniem" źródła danych odnośnie do systemu integracji danych. Zaproponowany model pozwalana na naturalne rozszerzenia, jak branie pod uwagę orzeczeń o jakości danych, a nawet bardziej wyraziste ontologie dla oceny jakości danych.

**Tematy** Jakość i oczyszczanie danych.

Jakość danych jest zależna od kontekStu

# Wstęp

Oszacowanie jakości źródeł danych jest zależne od kontekstu, tj. pojęcia „dobrych” lub „złych” danych nie mogą być oddzielone od kontekstu, w jakim dane są wytwarzane lub stosowane. Na przykład, dane o rocznej sprzedaży danego produktu z wahaniami sezonowymi mogą być uznane za dane dotyczące jakości przez analityka biznesowego oszacowującego roczny przychód produktu. Jednakże, te same dane mogą nie być wystarczająco dobre dla kierownika magazynu, który stara się oszacować zamówienia na następny miesiąc.

Ponadto, jakość danych jest związana z rozbieżnością pomiędzy rzeczywistymi przechowywanymi wartościami i "prawdziwymi" wartościami, które miały być lub oczekiwane jest że mają być przechowywane. Na przykład, jeżeli pomiar temperatury jest dokonywany z użyciem wadliwego termometru, zapisana wartość (pomiar) różniłaby się od właściwej wartości (rzeczywista temperatura), która była rzekomo tą jedną która miała być przechowywana. Jest to przykład *semantycznie nieprawidłowych danych* [3].

Ponadto, inny rodzaj semantycznej rozbieżności występuje wtedy, gdy *zmysły lub  
znaczenie* przypisywane przez różne środki do rzeczywistych wartości w bazie danych  
nie zgadzając się [19], jak pokazano w przykładzie 1. W tym artykule skupimy się na jakości danych (DQ – Data Quality) problemów powodowanych przez ten typ semantycznej rozbieżności.

*Przykład 1.* Tom jest pacjentem w szpitalu. Kilka razy na dzień jego temperatura jest mierzona i zapisywana przez pielęgniarkę. Jego lekarz, John chce zobaczyć temperaturę Toma codziennie około południa, aby śledzić jej zmiany. Informacja, której John potrzebuje pojawia się w relacji TempNoon w Tabeli 1, która zawiera temperatury pomiędzy 11:30 a 12:30 w ciągu dnia dla każdego z pacjentów Johna.

**TempNoon**

|  |  |  |  |
| --- | --- | --- | --- |
| **Pacjent** | **Wart.** | **Czas** | **Data** |
| Tom Waits | 38.5 | 11:45 | Sep/5 |
| Tom Waits | 38.2 | 12:10 | Sep/5 |
| Tom Waits | 38.1 | 11:50 | Sep/6 |
| Tom Waits | 38.0 | 12:15 | Sep/6 |
| Tom Waits | 37.9 | 12:15 | Sep/7 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Pacjent** | **Wart.** | **Czas** | **Data** |
| Tom Waits | 38.5 | 11:45 | Sep/5 |
| Tom Waits | 38.0 | 12:15 | Sep/6 |
| Tom Waits | 37.9 | 12:15 | Sep/7 |

John ma dodatkowe wymagania co do *jakości* pomiaru temperatury jego pacjentów: muszą one być przeprowadzane przez certyfikowane pielęgniarki z termometrem doustnym.5 września, nieświadoma nowych wymagań, Cathy mierzy temperaturę Toma o 12:10 za pomocą termometru *na podczerwień* i zapisuje wynik jako liczbę (krotkę) 2 w tabeli 1. Ponieważ wykorzystywany przyrząd nie pojawia się w widoku, który John ma w danych, interpretuje on wartość 38.2oC jako zmierzoną za pomocą termometru doustnego.

Jest to przykład rozbieżności pomiędzy semantyką wartości w sposób przewidziany przez producenta danych (38.2oC zmierzone za pomocą termometru na podczerwień) i semantyki oczekiwanej przez konsumenta danych (38.2oC zrobione za pomocą termometru doustnego). Ta krotka nie powinna pojawiać się w tabeli jakości, to znaczy w takiej, który spełnia wymagania Johna dotyczące jakości, ponieważ taka tabela będzie zawierać tylko temperatury zmierzone za pomocą termometru doustnego.

Podobny problem pojawia się w trzeciej krotce w Tabeli 1: Została zmierzona przez nową pielęgniarkę, Helen, która jeszcze nie jest certyfikowana, a tym samym nie spełnia jednego z wymagań lekarza. Ta krotka nie powinna pojawiać się w tabeli jakości zawierającej wyłącznie temperatury zmierzone przez dyplomowane pielęgniarki.

Tabela 2 rozwiązuje problemy związane z Tabelą 1 w odniesieniu do specyfikacji lekarza: Problematyczne druga i trzecia krotka nie pojawiają się w niej.

Jak możemy powiedzieć lub wierzyć że Tabela 2 zawiera wyłącznie dane jakościowe? Na pierwszy rzut oka nie różni się za wiele od Tabeli 1. Ta pozytywna ocena byłaby możliwa, jeśli mieliśmy bazę kontekstową zawierającą dodatkowe informacje, np. Tabele 3, 4 i 5.

|  |  |  |
| --- | --- | --- |
| **Data** | **Zmiana** | **Pielęg.** |
| Sep/5 | morning | Susan |
| Sep/5 | afternoon | Cathy |
| Sep/5 | night | Joan |
| Sep/6 | morning | Helen |
| Sep/6 | afternoon | Cathy |
| Sep/6 | night | Cathy |
| Sep/7 | morning | Susan |
| Sep/7 | afternoon | Susan |
| Sep/7 | night | Joan |

|  |  |  |
| --- | --- | --- |
| **Pielęg.** | **Data** | **Typ** |
| Susan | Sep/5 | Oral |
| Cathy | Sep/5 | Tymp |
| Joan | Sep/5 | Tymp |
| Helen | Sep/6 | Oral |
| Cathy | Sep/6 | Oral |
| Susan | Sep/7 | Oral |
| Joan | Sep/7 | Oral |

|  |  |
| --- | --- |
| **Imię** | **Rok** |
| Ann | 2003 |
| Cathy | 2009 |
| Irene | 2000 |
| Nancy | 1995 |
| Susan | 1996 |

Pierwsza relacja zawiera imię pielęgniarek na oddziale Toma Waitsa i zmiany na których pracują w ciągu dnia. Są to pielęgniarki robiące pomiary; ponieważ jest to mały oddział jest tylko jedna pielęgniarka na zmianę z tym zadaniem. Druga relacja rejestruje imię certyfikowanych pielęgniarek na oddziale, oraz rok w którym dostała certyfikat. Ostatnia relacja zawiera typ termometru którego każda pielęgniarka używa w ciągu dnia (na przykład doustny lub na podczerwień); każda pielęgniarka wykonuje wszystkie pomiary temperatury w ciągu dnia za pomocą tego samego rodzaju termometru. Ta informacja kontekstowa pozwala nam ocenić jakość danych w Tabelach 1 i 2.

W tym artykule bierzemy na poważnie intuicję i doświadczenie, że jakość danych jest zależna od kontekstu. Nasze sformalizowanie kontekstu podawane jest jako system zintegrowanych danych i metadanych, którego źródłem danych na podstawie oceny jakości jest szczególny i specjalny składnik. Dokładniej, kontekst dla oceny jakości danych w określonej instancji D schematu S jest podany przez wystąpienie I możliwie różnych schematów C, które mogłyby stanowić rozszerzenie S. W celu oceny jakości D, należy to "umieścić w kontekście", który został osiągnięty przez mapowanie D (oraz S) do kontekstowego schematu i danych. Właściwie, C może być bardziej skomplikowane, że jeden schemat lub instancja, a mianowicie zbiór schematów baz danych i instancji powiązanych ze sobą przez mapowania danych i schematu.

W naszym kontekście jakość instancji bazy danych D może być postrzegana jako "ślad" z kontekstowej, rozszerzonej bazy I. Ewentualnie dodatkowe informacje w I są tym co daje kontekst oraz wyjaśnia dane w schemacie D. Kontekstowy schemat i dane nie służą w danym przypadku do narzucenia jakości danych. Zamiast tego stosuje się: (A) Ocenianie jakości danych w instancji pod ręką; (B) Charakteryzowanie jakości odpowiedzi na pytania; oraz (c) Możliwość uzyskania tych odpowiedzi jakościowych do zapytania użytkownika. Wszystko to jest osiągnięte przez porównanie danej instancji D z instancją I, wirtualnej lub istniejącej, która może być zdefiniowana dla schematu kontekstowego na podstawie D źródeł zewnętrznych, które współpracują z danymi w schemacie kontekstowym i ewentualnych dodatkowych danych na poziomie kontekstowym, jak pokazano w przykładzie 1.

Powyższa instancja I może zostać zastąpiona przez znacznie bogatszy opis kontekstowy, na przykład pełnoprawnej ontologii. W tym stylu, ale ciągle w klasycznym scenariuszu bazy danych, możemy zdefiniować kilka dodatkowych *predykatów jakościowych* na C [19]. Mogą one być wykorzystane do oceny jakości danych w D (a także do jakości odpowiedzi zapytań z D, będziemy badać to później).

Następujące wkłady można znaleźć w tym artykule: (a) model kontekście oceny jakości danych. (b) jego stosowanie do czyszczenia lub odpowiadania na zapytania dotyczące jakości danych. (c) Jego zastosowanie do oceny jakości danych poprzez kilka naturalnych pomiarów, które powstają się z modelu. (d) Niektóre algorytmy dla wyżej wymienionych zadań w kilku konkretnych, ale wspólnych i naturalnych przypadkach. (e) stworzenie ram, które mogą być naturalnie rozszerzane w późniejszej pracy żeby objąć więcej ogólnej ontologii kontekstowych i zewnętrznie zdefiniowanych predykatów jakościowych.

Pozostała część artykułu jest zorganizowana w następujący sposób. W rozdziale 2 przedstawiono ogólne ramy kontekstowej jakości danych i zilustrowanie ich za pomocą działającego przykładu. W rozdziale 3, rozważamy dwa szczególne przypadki ogólnych ram, w których zakładamy, że mamy kontekstową instancję I, którą możemy wykorzystać do oceny jakości i przedstawić algorytm odpowiadający na zapytania dotyczące jakości w ramach tego założenia. W rozdziale 4, badamy bardziej skomplikowane przypadki, na przykład, gdzie nie istnieją takie przypadki kontekstowe. Omówimy podobne prace w pkt 5; i podsumujemy i odniesiemy się do na naszej bieżącej i przyszłej pracy w punkcie 6.

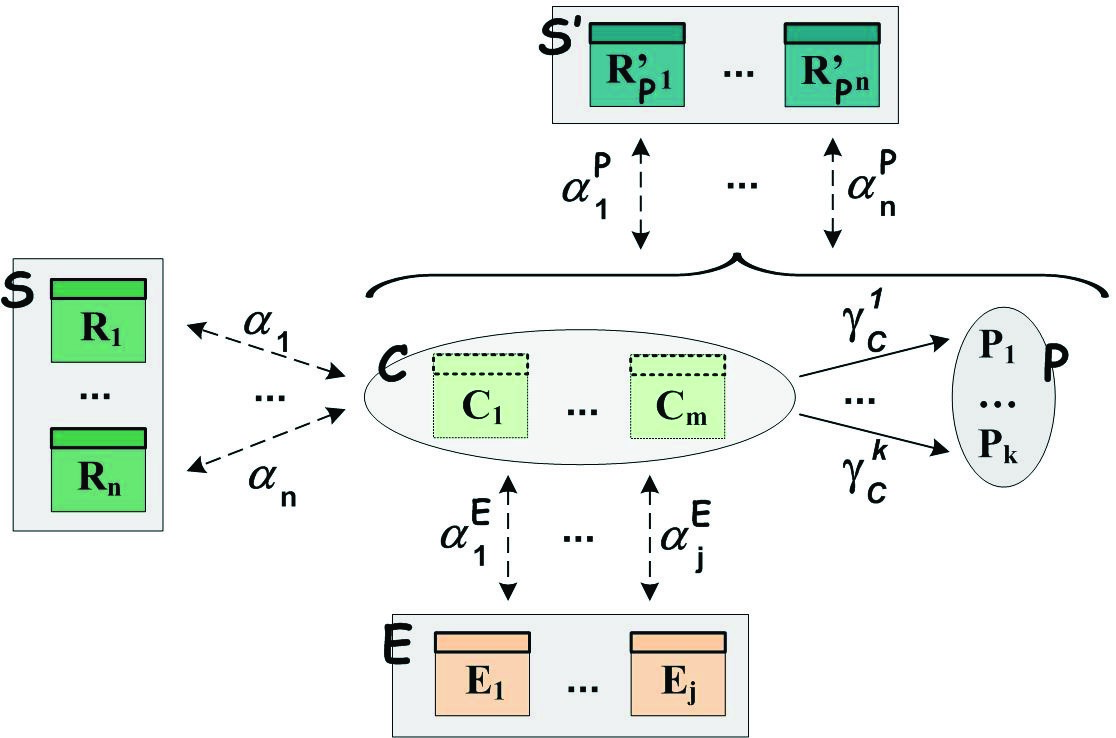
# Ramy dla jakości danych kontekstowych

Rozważmy relacyjny schemat S, z relacyjnymi predykatami *R,...* ∈ S. Ten Schemat określa język *L*(S) pierwszego rzędu logiki predykatów. W tym dokumencie bierzemy pod uwagę tylko zapytania jednostajne i widoki, na przykład koniunkcyjnych zapytań i ich związków, które zwykle piszemy w nie-rekursywnym Datalogu z wbudowaniami [1]. Rozważymy również instancję *D* z S, z rozszerzeniem *R*(*D*) dla R. itp. Jeśli instancje bazy danych są pomyślane jako skończone zbiory cząstek podstawowych, a następnie dla każdego *R* ∈ S, *R*(*D*) ⊆ *D*. Przypadki *D,R*(*D*)*,...* są to te podlegające oceny jakości w ramach systemu kontekstowego.

W najprostszej postaci, system składa się z kontekstowego relacyjnego schematu C, który może zawierać zbiór B wbudowanych predykatów. Możemy mieć lub nie instancję dla C. W bardziej skomplikowanym scenariuszu system kontekstowy może składać się z kilku schematów kontekstowych C1*,...,*C*n* a także zestawu zewnętrznych schematów E, które mogą być wykorzystywane przez system kontekstowy dla ocena instancji *D* od S.

W tych ogólnych ramach, uczestniczące schematy są powiązane przez *schemat mapowań*, podobnie jak te znajdujące się na przykład w systemach wirtualnych integracji danych (VDISs) [22, 4] lub wymiany danych [21], a nawet bardziej skomplikowane relacje logiczne, takie jak te, powszechne w systemach zarządzania danymi rówieśniczych [5, 6]. (Por [13] na podstawie analizy połączeń między tymi trzema obszarami.) Mapowania schematu przybierają formę korespondencji pomiędzy dwoma wzorami, jak zapytania lub widoki definicji, każdy z nich zawierające predykaty z jednego lub kilku schematów. W szczególności źródłem danych na podstawie oceny D będą mapowania do schematu kontekstowego. Powszechną formą połączenia lub mapowania jest z ∀*x*¯(*S*(¯*x*) → *φG* (¯*x*)), gdzie S jest relacyjnym predykatem źródła danych i *φ* (¯*x*) jest koniunkcyjnym zapytaniem na globalnym relacyjnym schemacie G. Te połączenie można znaleźć w *widokach lokalnych* (LAV) VDISs z otwartymi (lub dźwiękowymi) źródłami. Inną popularną formą połączenia jest forma ∀*x*¯(*ψ* (¯*x*) → *G*(¯*x*)), znajdujące się w global-as-view (GAV) VDISs ze źródłami otwartymi, gdzie *ψ* (¯*x*) jest koniunkcyjne zapytanie nad unią R z relacyjnych schematów u źródeł, i G jest globalnym relacyjnym predykatem. W *global-and-local-as-view* (GLAV) VDISs z otwartymi źródłami, znajdujemy skojarzenia między widokami (lub zapytaniami), z postaci ∀*x*¯(*ψ* (¯*x*) → *φ* (¯*x*)).

Rysunek 1 przedstawia ten ogólny scenariusz. Stosunki *Ri* w D podlegają ocenie jakości, która jest wykonywana za pomocą kontekstowego schematu C, która ma relacyjne predykaty *C*1*,...,Cm*. Istnieje również zbiór P *kontekstowych jakościowych predykatów* (CQPs) *P*1*,...,Pk*,, które są definiowane jako widoki C plus możliwie zewnętrznych źródeł *E*1*,...,Ej*.. W niektórych przypadkach, kombinacja schematów C, P, E może być postrzegana jako pojedynczy rozszerzony schemat kontekstowy. W innych przypadkach, to może być użyteczne, aby odróżnić je od siebie. Zewnętrzne predykaty *Ei* mogłyby mieć "pseudonimy" *Ei*′ w C, za pomocą prostych definicji jako mapowania w postaci ∀*x*¯(*E*′(¯*x*) ≡ *E*(¯*x*)) (lub w notacji Datalog, *E*′(¯*x*) ← *E*(¯*x*)).



**Rys. 1.** Ogólne ramy dla jakości danych kontekstowych

Z reguły będziemy mieć kopię S′ schematu S, z relacyjnymi predykatami , Chodzi o to, że *Ri*′ oznacza wersję jakości *Ri*, i ich rozszerzenia mogą być porównywane. Te idealne predykaty są powiązane ze (rozszerzonymi) kontekstowymi schematami przez mapowania *α*. Jeśli mapowanie dotyczy jakości predykatów P, zazwyczaj używamy predykatów *R*P′*i* i mapowań *α*P aby podkreślić zależność od P. W zasadzie każde z odwzorowań na Rysunku 1 może być dowolnego rodzaju LAV, GAV, GLAV plus dodatkowe założenia dotyczące otwartości/zamkniętości biorących udział instancji lub zwykłej definicji widoku.

Na przykład, *γ* są definicjami widoków predykatów jakościowych w zakresie elementów w C, plus możliwych źródeł zewnętrznych w E. Wszystkie te elementy, tak jak w przypadku integracji danych wirtualnych i systemów zarządzania danymi rówieśniczymi, określa zbiór dopuszczalnych instancji I do kontekstowej instancji C.

W kolejnych rozdziałach możemy rozważyć i zbadać kilka istotnych szczególnych przypadków tych ogólnych ram. W każdym z nich, zwracamy uwagę na: (a) Problem oceny jakości instancji D złożonej z relacji *R*1(*D*)*,...,Rn*(*D*). Ma to związek z analizą, jak różnią się one od idealnych, jakościowych instancji *Ri*. (b) Problem charakteryzowania i uzyskiwania wysokiej jakości odpowiedzi na pytania, które według przewidywań będą odebrane przez instancję D, która jest w trakcie oceny.

Jak widać, te dwa problemy są związane ze sobą. Rzeczywiście, początkowy problem może być postrzegany jako szczególny przypadek tego ostatniego. Możemy przedstawić główne idee już w ogólnych ramach: Biorąc pod uwagę różne schematy, mapowania, a niektóre zmaterializowane relacje (np. *Ri*, *Ek*, niektóre (części) *Cj*) relacyjne predykaty  będą miały (możliwe że wirtualny) dopuszczalne rozszerzenia, powiedzmy *Ri*′(*I*). Jakość *Ri*(*D*) jest określana przez jej "odległości" do *Ri*′(*I*), np. przez liczność, |*Ri*(*D*) △ *Ri*′(*I*)|, różnicy symetrycznej. Różne funkcje mogą być rozważane, zwłaszcza jeśli istnieje kilka dopuszczalnych rozszerzenia dla *Ri*′, jak to często ma miejsce w przypadku VDISs, gdzie może się pojawić kilka wirtualnych instancji *legalnych*.