5. Powiązane prace

Badania nad jakością danych obejmują zarówno charakteryzację typów błędów, modelowanie procesów, w których mogą one powstawać jak również opracowywanie sposobów ich wykrywania oraz eliminacji. Większość podejść do powyższych problemów, bazuje na domniemanym założeniu, że błędy w danych pojawiają się szczególnie na poziomie symboliczno-syntaktycznym, tzn. rozbieżności pomiędzy wartościami (przykładowo Kelvin, a Kelvn).

Jak pokazano w [19], problemy z jakością danych pojawiają się także na poziomie semantycznym, tzn. jako rozbieżności pomiędzy nadawanymi im znaczeniami i interpretacjami. Dokładniej, nawiązując do [19], problemy takie mogą pojawić się w przypadku różnicy między zamierzonym znaczeniem (przypisanym przez producenta danych) i interpretacją ich znaczenia (dokonaną przez konsumenta danych). Taka rozbieżność często spowodowana jest niejednoznacznością w komunikacji między producentem, a konsumentem danych. Taka wieloznaczność jest nieunikniona, gdy zmienne czynniki (jak typ użytego termometru lub stan pacjenta) nie są wyraźnie odnotowane w danych (bądź metadanych). Oczywiście to, czy taka wieloznaczność jest uznawana za problem z jakością danych, zależy od celu w jakim dane te są używane.

W [19] przedstawiono zarys metod definiowania zarówno syntaktycznej, jak i semantycznej jakości danych w jednolity sposób, opartych na podstawowych pojęciach wartości i znaczeń. Zostało również przedstawionych wiele wysokopoziomowych predykatów jakości danych, opartych na porównywaniu symboli i znaczeń (ścisłe dopasowanie, częściowe lub jego brak). W niniejszej pracy podejmujemy kroki, mające na celu zaproponowanie konkretnego mechanizmu zdobywania i porównywania na poziomie semantycznym wymagań dotyczących jakości danych, korzystając z kontekstowych relacji i predykatów jakości oraz pokazania ich zastosowań w realizacji zapytań.

Użycie kontekstów w zarządzaniu danymi było proponowane już wcześniej (w [9] znajdują się badania), jednakże wykorzystywało inne sposoby zdobywania, reprezentowania oraz wykorzystania kontekstów. Przykładowo, kontekstowe informacje były wykorzystywane do wspierania półautomatycznych procesów projektowania widoków (opisane w [8]). Kontekst w [8] składa się z wielu elementów będących parami klucz-wartość (np. stanowisko=”przedstawiciel”, sytuacja=”na miejscu”, czas=”dziś”). Pewne graniczenia również mogą być przedstawiane w kontekście (przykładowo gdy stanowisko to „menadżer”, sytuacja nie może odbywać się „na miejscu”).

Określenie kontekstu pozwala na wybór małego widoku spośród dużej struktury bazy danych, który jest uznawany za trafny w tym kontekście. Projektowanie widoku będącego świadomym kontekstu, gdy kontekst jest doprecyzowany może odbywać się ręcznie lub półautomatycznie poprzez łączenie częściowych widoków zależnych od poszczególnych elementów kontekstu [10]. W niniejszej pracy kontekst jest określony w podobny sposób, jednak w innym celu niż w [8]. Głównym celem [8] jest redukcja rozmiarów (np.: wydzielenie użytecznych danych w danym kontekście), natomiast niniejsza praca jako główny cel stawia wybór danych w zależności od ich jakości (np.: znalezienie takiego podzbioru danych, który najlepiej spełnia pewne wymagania jakościowe).

Pracą najbardziej powiązaną z szacowaniem jakości danych jest [24]. Propozycja Naumann’a bazuje na uniwersalnej relacji [23] utworzonej z globalnego, relacyjnego schematu do integracji niezależnych źródeł danych. Zapytania są zbiorem atrybutów uniwersalnej relacji, z możliwymi warunkami nałożonymi na wartości atrybutów. Aby odwzorować zapytanie w źródłowych widokach, zapytania użytkownika są tłumaczone na zapytania do globalnego schematu bazy danych. Naumann wyróżnia kilka kryteriów jakości źródeł danych, takich jak ich wiarygodność, obiektywność, reputacja oraz weryfikowalność. Kryteria te, są wykorzystywane przy określaniu modelu jakości dla zapytań.

Zgodnie z [24], jakość zapytania jest określana następująco. Każde źródło otrzymuje punktację jakości informacji (IQ – information quality) dla każdego kryterium uznawanego za istotne. Punktacje te są następnie formowane w wektor, gdzie każdy element odpowiada innemu kryterium. Użytkownicy mogą wyrazić swoje preferencje dotyczące danego kryterium poprzez nadawanie wag poszczególnym elementom wektora, tworząc wektor ważony. Wektor ten jest wykorzystywany z kolei przez wieloatrybutowe metody podejmowania decyzji (MADM multi-attribute decision-making) w celu utworzenia rankingu źródeł danych wchodzących w skład uniwersalnej relacji. Metodami tymi może być zarówno proste skalowanie i sumowanie punktacji jak i złożone wzory oparte o macierze zgodności. Model jakości nie zależy od wybranych metod MADM, jeżeli korzysta z wag użytkownika i punktacji jakości IQ. Celem jest, biorąc pod uwagę wektory IQ źródeł, uzyskanie wektora IQ planu zapytania zawierającego źródła. Plan zapytania może być rozumiany jako drzewo połączeń między źródłami: liśćmi są źródła danych, natomiast pozostałymi węzłami są złączenia. Punktacje jakości informacji są obliczane dla każdego wewnętrznego węzła, idąc od liści ku korzeniowi. Ogólna jakość planu zapytania dana jest w postaci punktacji jakości korzenia tego drzewa.

Ostatni akapit na 14