prof. UAM dr hab. Jerzy Szymański
Wydział Matematyki i Informatyki
Uniwersytet im. Adama Mickiewicza w Poznaniu

Poznań, 7 września 2018 roku

Recenzja pracy doktorskiej mgra Tomasza Piłki pt. „Eliminowanie redundancji i duplikatów w danych XML"

1. Omówienie zawartości pracy

Praca doktorska mgra Tomasza Piłki dotyczy zagadnień związanych z nadmiarowością danych w bazach XML-owych, liczy 133 strony i składa się ze wstępu, ośmiu rozdziałów merytorycznych, podsumowania, dodatku oraz spisu literatury, który obejmuje 110 pozycji (w tym sześciu, których mgr Tomasz Piłka jest autorem lub współautorem). Praca uzupełniona jest płytą CD zawierającą testową aplikację.

Rozdział 1 „Wstęp" jest krótkim przewodnikiem po zawartości całej pracy. Wprowadza nieformalne definicje oraz przedstawia cel pracy i zapowiada otrzymane wyniki.

W rozdziale 2 zatytułowanym „Relacyjny model danych" Autor przedstawia podstawowe definicje związane z relacyjnymi bazami danych. Mimo, że pojęcia te są powszechnie znane na uwagę zasługuje pełna i spójna formalizacja przedstawianych pojęć. Przedstawione definicje są bardzo precyzyjne, starannie zredagowane i spójne zarówno logicznie, jak i stylowo.

Rozdział 3 zatytułowany „Redundancja w bazach danych" poświęcony jest przedstawieniu pojęcia redundancji w relacyjnych bazach danych. Omówione jest samo pojęcie wraz z przykładami oraz przedstawione klasyczne metody usuwania redundancji związane z normalizacją schematu bazy danych. Pojęcia te zostały formalnie przedstawione przy pomocy rachunku krotek.

W rozdziale 4 zatytułowanym „Duplikaty w bazach danych" poświęconym przedstawieniu problemu identyfikacji duplikatów w bazach danych oraz bazach XML omówiono

znane z literatury algorytmy wykrywania duplikatów i przedstawiono ich słabe strony oraz ograniczenia. Zdefiniowano również formalnie pojęcie scalania krotek w dwóch przypadkach: gdy wśród różniących się wartości zawsze jedną jest NULL oraz ogólniej jako największy kres dolny.

W rozdziale 5 zatytułowanym „Model danych XML" Autor przedstawił własny sposób zdefiniowania schematu XML poprzez określenie pewniej gramatyki opartej o standard DTD (*Document Type Definition*). Ograniczeniem modelu jest to, że wszystkie dane muszą być zgodne z tą gramatyką. Dokument XML definiowany jest jako drzewo, którego wierzchołki mają identyfikatory oraz opcjonalnie etykiety i krotki). Ponownie definicje są sformalizowane i zgodne stylowo z poprzednimi definicjami.

Rozdział 6 „Zależności funkcyjne XML i postać normalna danych XML" wprowadza pojęcie zależności funkcyjnej XML oraz pojęcie spełnienia takiej zależności. Pokazuje również, że pojęcie spełnienia zależności funkcyjnej XML można traktować jako uogólnienie pojęcia spełnienia zależności funkcyjnej w relacyjnych bazach danych. Autor definiuje również pojęcie postaci normalnej XML oraz pokazuje, że dana XML, która nie jest w postaci normalnej zawiera redundancję.

W rozdziale 7 zatytułowanym „Informacyjne ujęcie redundancji danych" Autor wprowadza pojęcie miary ilości redundancji w danych XML. Miara ta zdefiniowana jest poprzez pojęcie entropii. Wszystkie pojęcia i definicje w tym rozdziale są zapisane bardzo formalnie i precyzyjnie.

Rozdział 8 zatytułowany „Eliminacja redundancji danych XML poprzez ich normalizację" zawiera najważniejsze wyniki teoretyczne tej pracy. Autor opracował metodę sprawdzania, kiedy algorytm normalizacji dla danych XML powinien być uruchomiony. W tym celu dowodzi twierdzenie podające warunek konieczny i dostateczny na posiadanie przez przez schemat XML postaci normalnej. Pozwala ono na sprawdzenie na poziomie schematu XML czy potrzebny jest proces normalizacji.

W rozdziale 9 zatytułowanym „Eliminowanie duplikatów danych XML" autor dokonuje przeglądu algorytmów eliminacji duplikatów oraz proponuje własne rozwiązanie oparte na scalaniu poddrzew. Dowodzi również twierdzenie mówiące, że jego algorytm eliminowania duplikatów nie narusza schematu XML.

W krótkim podsumowaniu (rozdział 10) Autor pokazuje ważność i użyteczność omawianych wyników oraz zwraca uwagę na swój wkład w tę tematykę.

Pracę kończy dodatek A „Implementacja biblioteki MergeXML," który zawiera opis oprogramowania przygotowanego do testowania rezultatów otrzymanych w pracy. Pełna wersja tego oprogramowania dołączona jest na płycie CD.

2. Ocena pracy

Wyniki uzyskane przez Autora w rozprawie dotyczą wykrywania i usuwania redundancji i duplikatów w danych XML. Problematyka ta jest bardzo ważna zarówno z teoretycznego, jak i praktycznego punktu widzenia. Wkład mgra Tomasza Piłki w tę tematykę można podzielić na trzy części:

- zaproponowanie spójnej formalizacji pojęć związanych z redundancją danych oraz z duplikatami zarówno w relacyjnych bazach danych, jak i danych XML,
- udowodnienie szeregu twierdzeń (m.in. dotyczących istnienia postaci normalnej schematów XML),
- zdefiniowanie, zbadanie i przetestowanie algorytmów eliminacji duplikatów w danych XML.

Wszystkie trzy aspekty badań prowadzą do ciekawej teorii pełnej potencjalnych zastosowań i dającej podstawy do dalszego rozwoju.

Aby zdefiniować i przeanalizować powyższe zagadnienia Autor sprawnie posługuje się pojęciami i technikami zarówno matematycznymi (logika matematyczna, rachunek prawdopodobieństwa, teoria informacji), jak i informatycznymi (bazy danych, struktury danych, algorytmika). Jego rozważania są bardzo pomysłowe, a uzyskane wyniki dają istotne ulepszenia w stosunku do istniejących wcześniej rozwiązań. Uzasadnienie ich poprawności, efektywności i innych własności mimo, że wykorzystuje standardowe narzędzia, wymagało dużej wiedzy, pomysłowości i sprawności technicznej.

Czytając pracę odczuwa się brak porównania zaproponowanych algorytmów z innymi algorytmami rozwiązującymi te same zadania. Praktyczne przetestowanie tych algorytmów na tych samych zbiorach rzeczywistych danych pozwoliłoby na lepsze porównanie jakości i efektywności tych algorytmów.

Pewien niedosyt sprawiła zbytnia „skromność" Autora. Czytając pracę trudno dociec, które wyniki są własne. Można to poznać jedynie po braku powoływania się na autorów oraz z krótkich informacji we „Wstępie" i „Podsumowaniu". Ze swej natury praca doktorska powinna eksponować wyniki autora. Brakuje również informacji o dalszych losach prezentowanych wyników (np. pierwsze prace Autora dotyczące rozpatrywanych zagadnień ukazały się już 10 lat temu i brak informacji o ich dalszych losach).

Przechodząc do ogólnej oceny wyników zawartych w rozprawie pragnę stwierdzić, że uzyskując je Autor wykazał się nieprzeciętną pomysłowością, pracowitością, sprawnością

techniczną i uporem w przezwyciężaniu trudności technicznych. Mimo, ze praca powstawała w bardzo długim okresie czasu, to udało zachować się pewną spójność. Autor wykazał się dużą pomysłowością i umiejętnością rozwiązywania praktycznych problemów informatycznych, potrafił również do analizy swoich rozwiązań zastosować zaawansowany aparat matematyczny.

Mgr Tomasz Piłka opublikował wyniki prezentowane w pracy w sześciu pracach oraz przedstawiał je na kilku konferencjach.

Reasumując, wyniki uzyskane w recenzowanej pracy doktorskiej oceniam wysoko. Sądzę, że omawiana rozprawa doktorska stanowi istotny wkład w rozwój teorii baz danych.

3. Uwagi redakcyjne

Praca napisana jest bardzo staranie, zarówno pod względem logicznym, jak i redakcyjnym. Poza drobnymi literówkami nie wpływającymi na ocenę pracy, nie znalazłem żadnych istotnych błędów. Poniżej lista przykładowych literówek:

- str. 10, wiersz 11: jest „złączeniowe," powinno być „złączeniowe;"
- str. 13, wiersz 11: jest „rzędy," powinno być „rzędu;"
- str. 25, wiersz 15: jest „dla z końcowych," powinno być „dla danych z końcowych;"
- str. 41, wiersz 9: jest „opracowane," powinno być „opracowano;" wiersz 6 od dołu: jest „ operuję," powinno być „operuje;" poza tym rysunek na tej stronie jest nieczytelny;
- str. 54, wiersz 8: jest „widoczne," powinno być „widoczna;" wiersz 14: jest „sencie," powinno być „sensie;"
- str. 81, wiersz 17: brak numeru twierdzenia;
- str. 128, wiersz 4: jest „Tomasz Piłka, M.K." powinno być „Piłka, T. and Kotecki, M." (w zasadzi ewszędzie w spisie literatury zamiast „and" powinno być „i").

4. Konkluzja

Uważam, że przedstawiona do oceny rozprawa mgra Tomasza Piłki zatytułowana „Eliminowanie redundancji i duplikatów w danych XML" spełnia wszystkie ustawowe i zwyczajowe wymogi stawiane pracom doktorskim i może stanowić podstawę do nadania

mgrowi Tomaszowi Piłce stopnia naukowego doktora nauk matematycznych, w dziedzinie informatyka. W związku z tym wnoszę o dopuszczenie rozprawy doktorskiej mgra Tomasza Piłki do publicznej obrony.