prof. dr hab. inż. Krzysztof Goczyła, prof. zw. PG
Politechnika Gdańska
Wydział Elektroniki, Telekomunikacji i Informatyki
kris@eti.pg.edu.pl

Gdańsk, 28.01.2019

Recenzja rozprawy doktorskiej

mgra Tomasza Piłki

(Uniwersytet im. Adama Mickiewicza w Poznaniu, Wydział Matematyki i Informatyki)

zatytułowanej:

Eliminowanie redundancji i duplikatów w danych XML

Wstęp

Niniejsza recenzja dotyczy drugiej, poprawionej i uzupełnionej wersji rozprawy, przesłanej na mój adres w dn. 7.01.2019 r. Wersję pierwszą skierowałem do poprawy pismem z dn. 26.09.2018 r. skierowanym na ręce Dziekana Wydziału Matematyki i Informatyki UAM prof. dr. hab. Jerzego Kaczorowskiego.

1. Problem badawczy i jego znaczenie

Temat rozprawy dotyczy problemu eliminowania informacji nadmiarowych (redundantnych) z danych zapisanych w formacie XML. Od początku bieżącego stulecia, kiedy to został sformułowany pierwszy standard XML, format ten, zwany też z uwagi na swoją elastyczność formatem semistrukturalnym, zdobył i nadal zdobywa liczne zastosowania. Wynika to z kilku przesłanek. Po pierwsze, format XML jest formatem tekstowym, co sprzyja wymianie informacji pomiędzy repozytoriami i systemami baz danych, które wewnętrznie stosują własne formaty przechowywania danych. Po drugie, struktura dokumentów XML jest znacznie bardziej ogólna niż struktura klasycznych relacji w relacyjnych bazach danych, z definicji spełniających pierwszą postać normalną (1NF). Taka elastyczność danych XML umożliwia lepsze, bardziej naturalne modelowanie złożonych obiektów świata rzeczywistego. Po trzecie, język XML jest językiem generycznym, który może służyć jako podstawa do tworzenia innych języków opartych na znacznikach, definiowanych przez stosowny dokument strukturalny XML Schema. Języki te, takie jak np. RDF i OWL, są podstawą tzw. Semantycznego Internetu (Semantic Web) – idei wzbogacenia treści zamieszczanych w sieci WWW o samoobjaśniające się elementy semantyczne.

W obszarze relacyjnych baz danych problem redundancji jest bardzo dobrze rozpoznany zarówno z praktycznego, jak i teoretycznego punktu widzenia z uwagi na elegancką teorię matematyczną stojącą za tymi bazami, obejmującą teorię normalizacji. W literaturze raportowane są też narzędzia (bardziej lub mnie eksperymentalne) wspomagające analityka w normalizacji baz danych. W przypadku baz XML-owych problem eliminowania redundancji jest dużo bardziej złożony i znacznie mniej rozpoznany. Z uwagi na złożoność i różnorodność struktury, dokumenty XML mogą zawierać informacje powtarzające się, które nie są łatwe do zlokalizowania na podstawie ręcznej analizy schematu lub – co jeszcze bardziej oczywiste – analizy samego dokumentu.

Doktorant w swojej pracy podjął się zadania opracowania nowych metod analizy dokumentów XML pod kątem eliminacji redundancji, w tym danych zduplikowanych. W świetle przytoczonych

wyżej argumentów podjętą tematykę uważam za ważną i aktualną dla współczesnej informatyki, szczególnie dla jej działu zwanego przetwarzaniem danych, a podjęte przedsięwzięcie badawcze – za celowe i zarazem ambitne.

2. Cele i tezy rozprawy

W podrozdziale 1.2 rozprawy Autor definiuje cel i tezę rozprawy. Celem była analiza problemów związanych z występowaniem redundancji i duplikatów w bazach danych XML. Ogólna teza rozprawy głosząca, że wykrywanie i eliminowanie występujących w bazie danych XML redundancji i duplikatów poprawia jakość zarówno schematu, jak i stanu bazy danych XML, może wydawać się trywialna, jednak jest ona uszczegółowiona do dwóch tez pomocniczych, które precyzują podjęte cele badawcze. Pierwszym z nich było opracowanie efektywnej metody rozstrzygania, czy dokument XML jest w tzw. postaci normalnej PNX na podstawie schematu dokumentu, czyli bez konieczności analizy jego instancji. Drugim celem szczegółowym było opracowanie metody skutecznego eliminowania duplikatów ze znormalizowanych dokumentów XML.

Chciałbym w tym miejscu podkreślić, że wprawdzie obecność w bazie danych duplikatów sama w sobie jest przejawem redundancji, to jednak celowe jest, tak jak to uczynił Autor w tytule i tezach rozprawy, rozdzielenie tych dwóch pojęć, gdyż zupełnie inne podejścia i metody dotyczą procesu normalizacji, a zupełnie inne procesu eliminowania duplikatów.

3. Wkład Autora

Oryginalny dorobek i wkład w dziedzinę Autor zaprezentował w rozdziałach od 7. do 10. pracy. W rozdziale 7. zaproponowana została ilościowa charakterystyka redundancji danych XML, z wykorzystaniem pojęcia entropii, powiązanej z zależnościami funkcyjnymi ZFX. W rozdziale 8. Autor zaproponował autorską metodę normalizacji danych XML, eliminującą niedoskonałości znanej z literatury metody opartej na algorytmie dekompozycji. W rozdziale 9. Autor wprowadza nowe metody usuwania duplikatów z baz danych XML oparte na scalaniu poddrzew dokumentu XML, a w rozdziale 10. prezentuje stosowne algorytmy implementujące te metody.

Generalnie, układ rozprawy jest jasny i logiczny, a przytoczona bibliografia jest reprezentatywna. Jednak część dotycząca relacyjnych baz danych mogłaby być nieco krótsza, a zamiast tego można było bardziej skoncentrować się na eksperymentalnej walidacji metod zaproponowanych przez Autora na rzeczywistych danych XML.

Na uznanie zasługuje indywidualny wkład Autor w dziedzinę, udokumentowany w rozprawie. Za ten wkład uważam:

1. Opracowanie metody weryfikacji na poziomie schematu danych XML tego, czy konieczne jest przeprowadzenie procesu normalizacji. Jest to o tyle istotne, że schemat bazy danych XML jest wyspecyfikowany w jednym dokumencie, natomiast dane mogą znajdować się w wielu, często bardzo obszernych dokumentach. Istotnym osiągnięciem Autora jest sformułowanie i udowodnienie twierdzenia o postaci normalnej PNX schematu XML. Przybliża ono analizę schematów XML do analizy schematów relacji w relacyjnych bazach danych. Warto podkreślić, że zaproponowana w rozprawie metoda uwzględnia możliwość zachowania wszystkich zależności funkcyjnych w danych XML, w odróżnieniu od procesów normalizacji relacyjnych baz danych, które mogą takich zależności nie zachowywać.

2. Opracowanie autorskiej metody eliminowania duplikatów w danych XML reprezentowanych w tzw. formie poszatkowanej. W ramach tej metody zaprezentowano i zaimplementowano algorytmy eliminacji duplikatów poprzez scalanie poddrzew dokumentu XML. Sformułowano też stosowne twierdzenie dotyczące jednego z algorytmów, choć do tego twierdzenia mam zastrzeżenia sformułowane poniżej, w punkcie 4. niniejszej recenzji.

3. Ujęcie schematu XML w postaci gramatyki, a dokumentów XML jako drzew zgodnych z tą gramatyką. Ważne jest to, że ta gramatyka posłużyła Autorowi do przeprowadzonej w rozprawie analizy redundancji i duplikatów w danych XML i do uzyskania oryginalnych wyników.

4. Dokonanie eksperymentalnej weryfikacji pewnego podzbioru zaproponowanych metod i przeprowadzenie dyskusji wyników z tych eksperymentów.

5. Dogłębne przeanalizowanie problemów związanych z normalizacją danych w relacyjnych bazach danych z odniesieniem do danych XML, a także dyskusja informacyjnego pojmowania redundancji danych opartego na entropii.

Podsumowując tę część recenzji, stwierdzam, że Autor rozprawy zrealizował postawione w rozprawie cele i uczynił to w sposób wyczerpujący i dojrzały z naukowego punktu widzenia. Tezy rozprawy można uznać za dostatecznie uzasadnione i zweryfikowane.

4. Ocena rozprawy i uwagi krytyczne

Rozprawa napisana jest w języku polskim. Zawiera 11 rozdziałów, bibliografię i dodatek dotyczący części implementacyjnej pracy. Układ rozprawy jest poprawny i logiczny, szkoda jednak, że Autor nie zamieścił na początku rozprawy spisu symboli, co ułatwiłoby śledzenie tekstu. Zamieszczona bibliografia jest obszerna – zawiera 110 pozycji z zakresu relacyjnych baz danych i danych XML-owych, a także zagadnień pokrewnych z zakresu przechowywania i przetwarzania informacji.

Czytając rozprawę, nasunęły mi się pewne uwagi o charakterze krytycznym, a także natrafiłem na pewne błędy. Prezentuję je poniżej.

1. Na str. 2 rozprawy Autor pisze (cytuję): „Ponieważ każda tabela (relacja) w relacyjnej bazie danych jest wielozbiorem (ang. *bag, multiset*), więc może zawierać dowolną liczbę wystąpień (egzemplarzy) każdej krotki”. Jest to stwierdzenie nieścisłe. Według teorii relacyjnych baz danych Codda, relacja bazodanowa jest matematycznym zbiorem, a więc z definicji nie zawiera duplikatów (powtarzających się krotek), w odróżnieniu od tzw. tabel tworzonych w konkretnej bazie danych odpowiednimi instrukcjami SQL. Niestety, Autor nie tylko w tym miejscu, ale także w innych miejscach rozprawy utożsamia te dwa pojęcia, co negatywnie wpływa na precyzję wywodów (na przykład we wstępie do rozdziału 2. czy też przy wprowadzeniu na str. 17 terminu „schematy relacyjne tabel” zamiast ścisłego „schematy relacji”).

2. W części praktycznej rozprawy Autor proponuje algorytm *UsunDup* eliminujący duplikaty z instancji danych XML, a następnie formułuje twierdzenie 10.2.1 dotyczące tego algorytmu. Twierdzenie to podane zostało bez dowodu, zapewne dlatego, że bardzo trudno byłoby je

udowodnić, tak jak jest trudno przeprowadzać dowody dotyczące nietrywialnych algorytmów. Myślę, że lepszym uzasadnieniem poprawności działania algorytmu *UsunDup* byłoby przeprowadzenie eksperymentów na bardziej złożonych danych niż to zaprezentowano w rozprawie. Można by też wówczas eksperymentalnie wnioskować o stopniu złożoności tego algorytmu, np. w funkcji liczby węzłów lub krawędzi drzewa dokumentu XML. W rozprawie nie znalazłem dyskusji na ten temat.

3. Twierdzenie 8.5.1 o weryfikacji postaci PNX, jeden z podstawowych wyników rozprawy, jest sformułowane tak, że podaje warunek dostateczny (występuje tam słowo „jeśli"). Tymczasem Autor dowodzi zarówno dostateczności, jak i konieczności. Sądzę zatem, że słowo „jeśli" powinno być zastąpione zwrotem „wtedy i tylko wtedy" lub równoważnym.

4. Na stronie 45. u góry podano oszacowanie złożoności pewnego algorytmu wyznaczania odległości Levenshteina. W tym oszacowaniu występuje symbol *max* z tylko jednym argumentem, co czyni tę formułę niepoprawną.

5. Przykład 6.1.1 dotyczy schematu *R* z rys. 5.4, a nie – jak błędnie podano – schematu *S*.

6. Rozprawa zawiera pewną liczbę błędów literowych, jednak ich liczba mieści się w granicach średniej. Poniżej przytaczam niektóre z nich, mające wpływ na znaczenie tekstu:
   - str. 3: *Nzwisko* zamiast *Nazwisko*,
   - str. 19: w definicji 2.2.3 jest „podzbioru" zamiast „podzbiorem",
   - str. 33: w wyrażeniu *Incl* występuje człon EgzStud[IdPrz] zamiast EgzStud[IdPrzed],
   - str. 43: pierwsze zdanie podrozdziału 4.3.1 wymaga przeformułowania,
   - Autor czasem pisze „Codda", a czasem „Codd'a"; poprawnie jest „Codda".

5. Podsumowanie

Pan mgr Tomasz Piłka w swojej rozprawie doktorskiej wykazał się bardzo dobrą znajomością problematyki przetwarzania danych, w szczególności problematyki konstrukcji i analizy baz danych XML. Przedstawił w sposób wyczerpujący aktualny stan wiedzy w zakresie eliminowania redundancji i duplikatów w systemach relacyjnych i systemach XML-owych, a dla osiągnięcia celów swojej pracy skorzystał ze znanych z literatury wyników, rozwijając je w sposób twórczy w celu wykazania prawdziwości postawionych w rozprawie tez. Istotne jest, że część swoich wyników zaimplementował i przetestował w postaci algorytmów komputerowych.

Stwierdzam, że w swojej rozprawie Doktorant wykazał się wiedzą i umiejętnościami wymaganymi do uzyskania stopnia doktora w dziedzinie **nauk matematycznych** w dyscyplinie **informatyka** zgodnie z *Ustawą z dn. 14 marca 2003 r. o stopniach naukowych i tytule naukowym* (z późn. zmianami). W konsekwencji wnoszę o przekazanie recenzowanej rozprawy do dalszych etapów przewodu doktorskiego.