

Prof. dr hab. Jarosław Stepaniuk

Katedra Systemów Informacyjnych i Sieci Komputerowych

Wydział Informatyki, Politechnika Białostocka

Recenzja Rozprawy Doktorskiej

Wybrane metody ewaluacji i strojenia bazodanowych silników aproksymacyjnych bazujących na granularnych podsumowaniach danych mgr Agnieszki Chądzyńskiej – Krasowskiej

Rozprawa nawiązuje do dziedziny przybliżonego przetwarzania zapytań (ang. Approximate Query Processing, w skrócie AQP) w systemach zarządzania dużymi bazami danych. Zamiast gromadzenia pełnego zbioru danych oraz dostarczenia użytkownikowi wszystkich możliwych krotek będących dokładnym wynikiem sformułowanego przez niego zapytania proponuje się w rozprawie wykonywanie zapytań na istotnie mniejszym w stosunku do całości danych zbiorze reprezentacji przybliżonych, które mogą przyjmować różne formy, od próbki pełnego zbioru danych począwszy, a na zwartym podsumowaniu skończywszy. Jest to ważne zagadnienie w informatyce i zajmuje się nim wielu badaczy zarówno w uczelniach zagranicznych jak i polskich.

Doktorantka sformułowała w rozprawie następujące dwie tezy:

„Teza 1. Możliwe jest zaprojektowanie rozszerzonego środowiska ewaluacyjnego dla silników bazodanowych typu AQP, uwzględniającego aspekt mierzenia dokładności wyników zapytań na podstawie podobieństwa między wynikami przybliżonymi i ich dokładnymi odpowiednikami, respektując praktyczne oczekiwania użytkowników odnośnie własności zdefiniowanej miary podobieństwa.

Teza 2. Możliwe jest wykorzystanie metod bazujących na podsumowaniach danych nie tylko na potrzeby klasycznych zapytań SQL, ale też w celu przybliżonego wykonywania zadań bardziej zaawansowanych, związanych np. z uczeniem maszynowym i wizualną analizą danych, działających w interakcji ze specjalnie zaprojektowanym repozytorium metadanych granularnych.”

Zagadnienie naukowe rozpatrzone w rozprawie jest sformułowane dostatecznie jasno przez Autorkę. Rozprawa ma charakter teoretyczny i doświadczalny – wyniki matematyczne rozprawy zostały uzupełnione prezentacją wyników badań eksperymentalnych.

Rozprawa została przygotowana w języku polskim i liczy 176 stron. Pokróćce przedstawię zawartość poszczególnych rozdziałów rozprawy.

Rozdział 1 pt. „Wstęp” zawiera cel, zakres i układ rozprawy. Rozdział 2 pt. „Zagadnienia wprowadzające” zawiera informacje dotyczące przybliżonego wykonywania zapytań. Omówiono też różne metody gromadzenia podsumowań. Rozdział 3 pt. „Dokładność wyników zapytań bazująca na podobieństwie” zawiera opis miar dokładności wyników zapytań analitycznych. Rozdział 4 pt. „Dostrajanie parametrów silnika granularnego” jest poświęcony opisowi rozważań nad dostrajaniem rozpatrywanego silnika do specyfiki danych i zapytań związanych z konkretnym wdrożeniem praktycznym. Jako przykładu użyto danych opisujących zdarzenia sieciowe oraz wymagań dotyczących wykrywania zagrożeń w ruchu sieciowym. Rozdział 5 pt. „Dostrajanie algorytmów podsumowujących” zawiera informacje na temat dostrajania algorytmów rozpatrywanego silnika bazującego na granularnych podsumowaniach danych. Rozdział 6 pt. „Granularne realizacje metod uczenia maszynowego” zawiera omówienie, między innymi, algorytmu mRMR (ang. minimum Redundancy Maximum Relevance), w którym kolejne atrybuty są iteracyjnie dodawane do zbioru wynikowego na podstawie siły zależności między danym atrybutem a atrybutem decyzyjnym oraz tymi atrybutami, które we wcześniejszych krokach weszły do zbioru wynikowego. W Rozdziale 7 pt. „Repozytorium metadanych granularnych” zaprezentowano propozycję repozytorium metadanych umożliwiającego łatwy dostęp do zgromadzonych podsumowań niezależnie od samego silnika. W ostatnim rozdziale Autorka zamieściła podsumowanie rozprawy. W dalszej części rozprawy zamieszczono liczącą 105 pozycji literaturę w porządku alfabetycznym. W uzupełnieniu zamieszczono dwa dodatki.

Układ rozprawy jest czytelny. Język i strona formalna rozprawy nie budzi większych zastrzeżeń. Drobne mankamenty stylistyczne są całkowicie usprawiedliwione biorąc pod uwagę złożoność omawianej tematyki jak i wielkość rozprawy.

Za główne wyniki rozprawy doktorskiej uważam:

- zaproponowanie podejścia ewaluacyjnego dla bazodanowych silników aproksymacyjnych, uwzględniającego aspekt dokładności obliczeń (wyrażanej za pomocą miary dokładności wyników zapytań analitycznych bazującej na

podobieństwie między uzyskiwanymi przy danych ustawieniach badanego silnika wynikami przybliżonymi i ich dokładnymi odpowiednikami),

- pokazanie, jak wykorzystać zaproponowane podejście do oceny dokładności zapytań wykonywanych w sposób przybliżony oraz przeprowadzenie szeregu testów porównawczych, które doprowadziły do udoskonalenia algorytmów zaimplementowanych w przykładowym silniku aproksymacyjnym bazującym na granularnych podsumowaniach danych,
- wzbogacenie zaproponowanego podejścia ewaluacyjnego o funkcjonalności wykraczające poza język SQL, w celu bardziej wieloaspektowej oceny, na ile obliczenia bazujące na danych zapisanych w sposób przybliżony mogą odbiegać od analogicznych obliczeń bazujących na danych dokładnych, w tym, przygotowanie projektu repozytorium metadanych reprezentujących granularne podsumowania danych oraz opracowanie przykładowych nowych wersji algorytmów uczenia maszynowego działających na podsumowaniach i pokazanie, jak badać podobieństwo między wynikami tych algorytmów otrzymanymi dla ich wersji klasycznej i aproksymacyjnej.

Do zalet rozprawy doktorskiej zaliczam:

- rozprawa obejmuje interesujące i aktualnie uznawane, w informatyce, za istotne zagadnienia,
- problem badawczy, rozpatrywany w rozprawie, został klarownie zdefiniowany,
- podkreślono znaczenie włożonej pracy w rozwój nauki,
- pierwsze rozdziały zawierają ciekawe wprowadzenie do problematyki stanowiącej obszar badań oraz związane z tematem definicje i wyjaśnienie najważniejszych zagadnień,
- jasno sformułowany został zakres badań empirycznych i zastosowane metody,
- wiele z omawianych kwestii jest zilustrowanych przykładami,
- w interesujący sposób przedstawione zostały wyniki eksperymentalne.

Do mankamentów rozprawy doktorskiej zaliczam nadmierną długość rozprawy.

Jeżeli chodzi o opublikowane wyniki badań Autorki, to przykładowo w popularnej, cieszącej się uznaniem naukowców, bazie publikacji informatycznych DBLP (<https://dblp.uni->

trier.de/pers/hd/c/Chadzynska=Krasowska:Agnieszka) znajdują się informacje na temat sześciu prac współautorskich Doktorantki opublikowanych w sprawozdaniach konferencyjnych w latach 2016 – 2018.

Konkluzja

Oceniana rozprawa wymagała przeprowadzenia badań teoretycznych oraz badań eksperymentalnych. Autorka wykazała bardzo dobre opanowanie warsztatu badawczego. Dlatego też, stwierdzam, że oceniana rozprawa **spełnia wymagania** stawiane rozprawom doktorskim przez obowiązujące przepisy i wnoszę o dopuszczenie Pani Agnieszki Chądzyńskiej – Krasowskiej do dalszych faz przewodu doktorskiego.



Prof. Jarosław Stepaniuk