dr hab. inż. Sebastian Deorowicz,
profesor nadzwyczajny Politechniki Śląskiej
Instytut Informatyki
Wydział Automatyki, Elektroniki i Informatyki
Politechnika Śląska
44-100 Gliwice, ul. Akademicka 16

Gliwice, 8.10.2015

Recenzja rozprawy doktorskiej

Tytuł rozprawy:

Przeszukiwanie przestrzeni rozwiązań w optymalizacji planów zapytań do baz danych z wykorzystaniem heurystycznego algorytmu IWO

Autor:

mgr inż. Daniel Kostrzewa

Promotor:

Prof. dr hab. inż. Stanisław Kozielski

I. Problematyka naukowa oraz przedmiot rozprawy

W dzisiejszych czasach bazy danych wykorzystywane są w bardzo wielu zastosowaniach. Co więcej, rozmiary przetwarzanych danych nieustannie rosną. Stanowi to nie lada problem, z którym można walczyć inwestując w sprzęt oraz w oprogramowanie. Pierwsza możliwość jest dość oczywista, ale wymaga ponoszenia znaczących kosztów, zarówno zakupu, jak i utrzymania drogich serwerów. Drugą możliwością są inwestycje w jak najnowsze oprogramowanie zarządzające bazami danych. Jednym z istotnych kierunków rozwoju tego oprogramowania są optymalizacje realizacji zapytań, dzięki którym możliwe jest, często znaczne, zredukowanie rozmiaru wyników pośrednich, co wprost przekłada się na czas realizacji zapytań. Wszystko wskazuje na to, że w przyszłości oba te kierunki będą musiały być rozwijane niezależnie.

Recenzowana praca dotyczy właśnie problematyki optymalizacji złożonych zapytań, w których występuje wiele złączeń.

II. Analiza treści rozprawy oraz uzyskanych wyników

1. Treść rozprawy

Doktorant postawił sobie w pracy następujący cel:

- **Opracowanie zmodyfikowanej metody *Invasive Weed Optimization* dla potrzeb zadań optymalizacji, w szczególności wyznaczania kolejności złączeń w zadaniach optymalizacji zapytań do baz danych.**

W związku z tym celem, postawiono następującą tezę, którą Doktorant w swojej pracy dowodzi:

- **Zaproponowany w pracy zmodyfikowany algorytm *Invasive Weed Optimization* może zostać skutecznie wykorzystany w zadaniach optymalizacji, a w szczególności do wyznaczania kolejności złączeń w procesie realizacji zapytań w bazach danych.**

Rozprawa składa się z siedmiu rozdziałów, dwóch dodatków oraz bibliografii. Pierwsze cztery rozdziały mają charakter opisu stanu wiedzy.

Rozdział 1. jest krótkim wprowadzeniem, zawierającym streszczenie oraz określenie celu pracy.

Rozdział 2. przedstawia podstawowe pojęcia związane z optymalizacją zapytań w bazach danych.

W rozdziale 3. zawarto dokładniejsze omówienie problemu optymalizacji kolejności złączeń w zapytaniach bazodanowych.

Przegląd algorytmów optymalizacyjnych, ze szczególnym uwzględnieniem algorytmu IWO znajduje się w rozdziale czwartym.

Kolejne trzy rozdziały zawierają omówienie wyników własnych Autora.

I tak, w rozdziale 5. przedstawiono zmodyfikowany algorytm IWO. Autor wyznaczał tu m.in. złożoność jego obliczeniową. Ponadto pokazał jak zastosować ten algorytm dla trzech problemów optymalizacyjnych: znajdowania minimum funkcji wielowymiarowych, wyznaczania marszruty komiwojażera, określania kolejności złączeń w realizacji zapytań w scentralizowanych bazach danych.

Rozdział 6. zawiera omówienie wyników eksperymentów, w których Autor porównał zaproponowany przez siebie algorytm heurystyczny z algorytmami znanymi z literatury. Do porównania wykorzystano wiele różnych zestawów danych testowych.

Ostatni rozdział zawiera podsumowanie wyników rozprawy.

2. Najważniejsze wyniki przedstawione w rozprawie

Najważniejszym wynikiem rozprawy jest opracowany autorski wariant algorytmu optymalizacji IWO. Istotą działania oryginalnego algorytmu jest przeprowadzanie optymalizacji problemów ciągłych bądź dyskretnych poprzez wykorzystanie mechanizmu rozprzestrzeniania się chwastów w przyrodzie. Autor zaproponował kilka rozszerzeń oryginalnego algorytmu IWO. W szczególności, zaproponował dwa dodatkowe sposoby rozsiewania ziaren. Istotą tych modyfikacji jest wprowadzenie nowych sposobów eksploracji przestrzeni rozwiązań optymalizowanego problemu.

Po sformułowaniu zaproponowanych ulepszeń, Autor zweryfikował opracowany algorytm rozwiązując trzy klasyczne problemy optymalizacyjne. Warto podkreślić, że

wszystkie te problemy mają duże znaczenie praktyczne. W szczególności Doktorant pokazał jak zaadoptować ogólną ideę opracowanej metaheurystyki do rozwiązywania tych problemów. Następnie przeprowadził dość wyczerpujące badania eksperymentalne. Podjął także próbę wyznaczenia złożoności obliczeniowej opracowanej metody, zarówno w przypadku ogólnym, jak i w przypadku głównego problemu, którego dotyczy rozprawa, a więc optymalizacji kolejności złączeń w zapytaniach do bazy danych.

Uzyskane wyniki są wartościowe i mogą mieć istotne znaczenie praktyczne.

3. Uwagi merytoryczne

Rozprawa jest, w większości, napisana starannie. Autor starał się wyrażać w miarę precyzyjnie i formułować algorytmy w postaci pseudokodów, bądź schematów blokowych. W przeważającej części, opis algorytmów jest jasny i precyzyjny. Część eksperymentalna jest zadowalająca. Pozytywne wrażenie budzi także dość dużych rozmiarów (148 pozycji) wykaz cytowanej (adekwatnie) literatury.

Oceniając poprawność i oryginalność tezy należy stwierdzić, że została ona postawiona poprawnie, jest oryginalna i w rozprawie została wykazana.

Uzyskane przez Autora wyniki są ciekawe i na pewno zasługują na dalsze badania. Warto podkreślić, że porównanie z wysoce zoptymalizowanym serwerem bazy danych jakim jest MS SQL Server 2008 wypadło na korzyść Doktoranta, tzn. w ramach rozwiązań, które zaimplementował Autor, jego optymalizator pozwala na uzyskanie wyników zapytań szybciej niż optymalizator stosowany w komercyjnym systemie.

Autor wykazał się dość dobrą umiejętnością przedstawiania wyników przeprowadzonych przez siebie badań. W szczególności nie unika porównań z najnowszymi wynikami z literatury.

Generalnie moja ocena merytorycznej jest jak najbardziej pozytywna. Poniżej skupię się jednak na pewnych zauważonych uchybieniach.

4. Uwagi krytyczne i redakcyjne

Na tle dość klarownego opisu, zawierającego wartościowe wyniki, dość niefortunnie prezentuje się przeprowadzana w kilku miejscach rozprawy analiza złożoności obliczeniowej algorytmu IWO oraz jego zmodyfikowanej wersji. Poniżej odniosę się do ważniejszych z zauważonych w tym względzie uchybień i błędów.

1. Formułując wzory należało zapisać je starannie jako wzory, a nie jako teksty. W szczególności mam tu na myśli wykorzystanie symboli, które w sposób zwięzły opisywałyby parametry, funkcje. Zamiast tego Autor stosuje rozwlekłe opisy typu „liczba_chwastów (ze względu na inicjalizację)", co powoduje, że relatywnie proste wzory są zapisywane w wielu wierszach, niepotrzebnie utrudniając w ten sposób lekturę i zrozumienie sensu wyrażeń.

2. Dość niespotykane jest stosowanie komentarzy wewnątrz wzorów. Skutecznie powodowało to, że w początkowej fazie lektury nie byłem w stanie zrozumieć wzorów. Komentarze należy umieścić w tekście omawiając wyrażenie.

3. W literaturze zwykło się używać notacji $O()$ (nawiasy okrągłe) a nie $O[]$ (nawiasy kwadratowe).

4. Niejasno zdefiniowano czym są operacje dominujące w rozważanych problemach.

5. Wzór (5) na str. 24 jest niepoprawny. Składnik logarytmiczny opisuje koszt wstawienia pojedynczego elementu do struktury drzewiastej zamiast kosztu wstawienia wszystkich elementów. W konsekwencji niepoprawne są wzory (6) i (7).

6. Dość nieszczęśliwe jest sformułowanie „Zakładając, że liczba chwastów jest duża, można pominąć wyrażenie z logarytmem, którego wartość jest znacznie mniejsza". W sensie O-notacji zawsze wyrażenie $O(x+\log x)$ można (i należy) uprościć do $O(x)$. Dotyczy to m.in. wzorów (7), (19).

7. Niepoprawny jest wzór (17) — przez analogię do wzoru (5). W konsekwencji niepoprawne są kolejne wzory, w których wykorzystuje się wzór (17).

8. We wzorze (27) Autor podjął interesującą próbę uwzględnienia czasu operacji dyskowych. Problem jednak w tym, że jeśli w prawym składniku sumy uwzględniamy rozmiar rekordu, to nie można go pomijać w lewym składniku. W obecnej wersji wzoru lewy składnik jest poprawny przy założeniu, że przetwarzanie rekordu odbywa się w czasie stałym, co nie może być prawdą w sytuacji, w której rozmiar rekordu jest parametrem a długość słowa maszynowego procesora jest stała.

9. Czy we wzorze (30) wszystkie parametry powinny być mnożone przez siebie?

10. Z opisu nie wynika dlaczego we wzorze (34) jeden z czynników pojawia się w drugiej potędze.

11. Stwierdzenie, że z zestawienia zależności (42), (43), (45) wynika, że najwyższą złożonością charakteryzuje się technika staczania nie jest poprawne. Nie zauważyłem w pracy założenia dot. zależności pomiędzy liczbą generowanych sąsiadów a liczbą przejść, liczbą zbiorów pierwotnych i liczbą możliwych złączeń, które uzasadniałoby taki wniosek.

Inne uwagi merytoryczne do pracy:

12. Na str. 8 Autor przedefiniowuje znaczenie słowa „optymalny". Sugerowałbym powstrzymać się od tego kroku, ponieważ słowo to ma dobrze określone znaczenie, tymczasem jego nowa definicja jest niezbyt jasna.

13. Problem komiwojażera został prawdopodobnie po raz pierwszy zdefiniowany w latach 30. XIX w., a nie jak zaznaczono w pracy (str. 16) w XX w.

14. W problemie komiwojażera nakładanie dodatkowego założenia na punkt startowy jest bezzasadne (str. 16), ponieważ i tak szukamy cyklu.

15. Na str. 44 Autor wprowadza metodę wykrywania kształtu złączeń. Brak jest jednak wyprowadzenia wzoru oraz uzasadnienia, że wartość L jednoznacznie definiuje „kształt" grafu.

16. Autor omawiając różne metaheurystyki optymalizacyjne wskazuje do jakich problemów były one stosowane. Interesującym byłoby jednak kilka zdań komentarza z jakim skutkiem się to odbywało (np. które metaheurystyki dawały najlepsze rozwiązania).

17. Korzystne dla czytelności byłoby przedstawienie zaproponowanego algorytmu w postaci pseudokodu. Łatwiej można by wtedy dyskutować kwestie złożoności obliczeniowej.

18. Brak jest informacji jak wstępnie wyznaczano współczynniki algorytmu (np. strony 55, 64, 74).

19. Opis tabeli powinien zawierać informację o znaczeniu zawartych w niej informacji. Przykładowo w tabeli 5 znajduje się wiele liczb a dopiero z lektury tekstu dowiadujemy się co te wartości znaczą.

20. Przy omawianiu wyników Autor czasami zamiennie stosuje nazwy IWO oraz exIWO. Jest to o tyle mylące, że tylko algorytm exIWO jest propozycją autorską.

5. Uwagi redakcyjne

Rozprawa została napisana w większości starannie. Występują w niej nieliczne literówki, jednak pojawiają się one na tyle rzadko, że w żadnym razie nie są irytujące dla czytelnika.

6. Podsumowanie

Wymienione powyżej uwagi merytoryczne nie mają istotnego wpływu na wagę i jakość wyników uzyskanych przez Doktoranta i omówionych w recenzowanej rozprawie. Opracowany przez Doktoranta algorytm wraz z wynikami eksperymentów są moim zdaniem interesujące z naukowego punktu widzenia i dowodzą przyjętej w pracy tezie.

Liczba pozycji literaturowych jest w zupełności wystarczająca. Pozycje te zostały dobrane poprawnie i nie mam w tym zakresie uwag krytycznych. Cytowane pozycje świadczą o głębokiej wiedzy doktoranta z zakresu badanej dziedziny.

III. Konkluzja

Rozprawa doktorska mgr. inż. Daniela Kostrzewy zawiera oryginalne i interesujące wyniki naukowe dotyczące metod przyspieszania realizacji zapytań do baz danych.. Zawarte w recenzji uwagi krytyczne nie wpływają na moją ogólną bardzo dobrą ocenę rozprawy. Uważam, że zostały spełnione wymagania stawiane rozprawom doktorskim przez *Ustawę o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki*. Wnoszę zatem o dopuszczenie wspomnianej rozprawy do publicznej obrony.