prof. dr hab. inż. Aleksander Byrski
Instytut Informatyki
Akademia Górniczo-Hutnicza
im. Stanisława Staszica w Krakowie
Al. Mickiewicza 30, 30-059 Kraków
olekb@agh.edu.pl

Kraków 25.05.2022

Recenzja rozprawy doktorskiej pt. Developing efficient methods for mining inter-transaction patterns" opracowanej przez mgr inż. Thanh Ngo Nguyena, uczestnika Szkoły Doktorskiej Politechniki Wrocławskiej

Mgr inż. Thanh Ngo Nguyen (dalej "Doktorant") przedstawił do oceny rozprawę doktorską w formie manuskryptu opracowanego w języku angielskim wraz z ustawowo wymaganym streszczeniem w j. polskim, liczącego 152 strony o klasycznej strukturze pracy naukowej (wprowadzenie z podstawowymi definicjami i postawieniem tezy a także organizacją pracy i określeniem osiągnięć w niej przedstawionym, przegląd literatury wraz z identyfikacją aspektów, na których poprawienie ukierunkowana jest rozprawa, następnie przedstawia proponowane algorytmy eksploracji międzytransakcyjnych wzorców wraz z dyskusją badań eksperymentalnych, po tym rozdziale zamieszczona została konkluzja z określeniem przyszłych badań i ok 70 odniesień do bibliografii tematu).

Praca wpisuje się bardzo dobrze w obszar eksploracji danych na poziomie technicznym, mając na celu ulepszenie istniejących algorytmów eksploracji wzorców międzytransakcyjnych (znajdowanych w wielu transakcjach, oddzielonych czasem), aby były bardziej efektywne pod względem kosztu czasowego i pamięciowego.

Hipoteza sformułowana przez Doktoranta jest następująca: możliwa jest poprawa czasu wykonywania oraz wykorzystania pamięci dla procesu eksploracji danych międzytransakcyjnych dzięki wykorzystaniu kompaktowych struktur danych oraz efektywnych strategii pruningu. Sposób sformułowania hipotezy nie budzi zastrzeżeń i wpisuje się w dobre praktyki zastosowania metody naukowej.

Doktorant formułuje konkretne cele, których realizacja ma doprowadzić do wykazania prawdziwości postawionej hipotezy:

1. Opracowanie dwóch wydajnych metod eksploracji wzorców międzytransakcyjnych (FITPs)
    - Doktorant modyfikuje strukturę danych N-list zaproponowaną przez Deng et al 2012, tworząc strukturę ITN-list, która z założenia ma być odpowiednia do eksploracji FITPs przechowując informacje nt. wzorców-kandydatów. Następnie Doktorant wprowadza dwie strategie pruningu wzorców, które jednak nie spełniają kryterium bycia wzorcami "częstymi". Korzystając z cech listy ITN-list można również szybko obliczyć moc zbioru "support" dla ITPs. Doktorant wykorzystuje zbiór różnicowy diffset zamiast zbioru identyfikatorów tidset co ma pozytywny wpływ na zajętość pamięci. Doktorant opracowuje dwie efektywne strategie pruningu redukując również dzięki temu przestrzeń poszukiwań. Proponowane metody, DITP-Miner oraz NL-ITP-Miner

poprawiają testowane metody pod względem efektywności wykorzystania pamięci oraz czasu wykonania.

2. Opracowanie dwóch wydajnych metod eksploracji domkniętych wzorców międzytransakcyjnych (FCITPs).
    - Doktorant w tych badaniach wykorzystuje również listę ITN-list oraz diffset, proponując algorytm eksploracji FCITPs, również proponując dwie strategie pruningu w celu szybkiego ustalenia domkniętych cech ITPs. Algorytm NCITP-Miner znacznie redukuje przestrzeń poszukiwań i przyspiesza proces eksploracji. Algorytm wykorzystujący diffset - FCITP-Miner wykorzystuje dwie strategie pruningu, poszukując domkniętych wzorców korzystając z przeglądania drzewa wgłąb. Dwa proponowane algorytmy są porównywane z algorytmem ICMiner przy użyciu popularnych w środowisku eksploracji danych benchmarkowych baz danych. Podobnie jak w poprzednim punkcie, proponowane algorytmy charakteryzują się niższym kosztem pamięciowym oraz szybszym działaniem.
3. Opracowanie trzech wydajnych metod eksploracji maksymalnych wzorców międzytransakcyjnych (FMITPs).
    - Doktorant opracowuje trzy algorytmy eksploracji maksymalnych wzorców międzytransakcyjnych: nMITP-Miner, dMITP-Miner oraz tMITP-Miner, wykorzystując odpowiednio ITN-list, diffset oraz tidset,
    - Zaproponowano strategię pruningu w celu redukcji przestrzeni przeszukiwań, w większości przypadków nMITP-Miner charakteryzuje się niższym czasem wykonania w porównaniu dMITP-Miner oraz tMITP-Miner. Wykonane eksperymenty przy użyciu baz benchmarkowych wskazują na to że wszystkie trzy opracowane algorytmy są bardziej wydajne niż metody z literatury, z którymi są porównywane.

W tym miejscu należy zaznaczyć, że w zasadzie wszystkie opracowane przez Doktoranta algorytmy wykorzystują w sposób kreatywny wspomniane struktury danych, takie jak ITN-list, diffset czy tidset oraz dedykowane strategie pruningu, a ich efektywności dowodzą przedstawione i szeroko skomentowane eksperymenty. Takie podejście, zakładające poprawienie istniejących algorytmów, dokładnie przedstawiając co zostało poprawione i w jaki sposób wpłynęło to na zachowanie algorytmu (bazując na eksperymentach) znakomicie wpisuje się w stosowanie metody naukowej. Przedstawione wykresy w części eksperymentalnej są czytelne a komentarze sformułowane w sposób jasny, pozwalający sobie wyrobić jednoznaczne zdanie na temat przeprowadzonych eksperymentów i ich wyników.

Doktorant w pracy podaje szereg pseudokodów proponowanych algorytmów, które w sposób precyzyjny pokazują ich strukturę i działanie, wydaje się jednak, że dodanie rysunków przedstawiających działanie algorytmów w sposób schematyczny (nie chodzi mi tutaj o diagramy blokowe, tylko o pokazanie idei, bez żadnej konkretnej notacji), mogłoby służyć jeszcze lepszej prezentacji wprowadzonych innowacji.

Eksploracja jest procesem deterministycznym więc same algorytmy są deterministyczne, jednak wpływ na wykonanie złożonego eksperymentu w środowisku komputera może być obciążone niepewnością pomiaru, związaną z uruchomieniem (bądź nie) innych usług. Czy Doktorant sprawdził powtarzalność eksperymentów?

Doktorant w części eksperymentalnej skupił się na efektywności czasowej i pamięciowej, zupełnie pomijając porównanie algorytmów referencyjnych z nowatorskimi pod względem eksplorowanych reguł (zbiory reguł powinny być tożsame). Ze względu na charakter wprowadzonych ulepszeń, mających na celu poprawę efektywności, najprawdopodobniej zbiór reguł powinien być identyczny dla wszystkich wersji testowanych algorytmów, jednak byłbym wdzięczny za odniesienie się przez Doktoranta do tej sprawy na obronie. Obserwacja konkretnych wybranych przypadków mogłaby też pomóc w interpretowalności wyników - aktualnie bardzo ważnej z punktu widzenia eXplainable AI - z którą to dziedziną niewątpliwie praca Doktoranta jest związana bezpośrednio.

Obciążające zadania obliczeniowe i walka o ich efektywną implementacje stają się bardzo często przyczynkiem do podejmowania badań nad wykorzystaniem środowisk współbieżnych czy rozproszonych, szczególnie w dzisiejszych czasach gdy tak łatwy dostęp mamy choćby do infrastruktury typu GPGPU, a dla naukowców w Polsce również niezmiernie łatwy dostęp do infrastruktury superkomputerowej. Chętnie usłyszałbym od Doktoranta dyskusję potencjalnych możliwości rozwoju proponowanych algorytmów w taki sposób, aby zaimplementować je (choć częściowo) na GPGPU czy z wykorzystaniem HPC. Z pewnością nie jest to łatwa sprawa, gdyż poszukiwanie wzorców w danych łączy się z korzystaniem z wiedzy globalnej, a wszelkie uaktualnienia wiedzy globalnej bardzo mocno będą wpływać na skalowalność rozwiązania.

Z informacji zamieszczonej w pracy wynika, że Doktorant przeprowadzał swoje eksperymenty wykorzystując komputer klasy PC, tym bardziej interesuje mnie możliwość uogólnienia implementacji i przyspieszenia jej przy użyciu dostępnego sprzętu. Dodatkowo chciałem jeszcze stwierdzić, że obciążające operacje analityczne są często implementowane w bazach danych z wykorzystaniem funkcji i procedur składowanych. Chętnie usłyszałbym co Doktorant sądzi o możliwości rozwinięciu swoich algorytmów w taką stronę aby zostały zaimplementowane wewnątrz silnika bazodanowego, wyzwalane w czasie dodawania nowych transakcji, działające na dedykowanych, dodatkowych tablicach. Implementacja wszelkich algorytmów powinna być możliwa ze względu na to, że od wielu lat dialekty SQL (choćby T-SQL z którym Doktorant prawdopodobnie się spotkał implementując swoje algorytmy z użyciem Visual Studio) są kompletne w sensie Turinga.

Uwaga techniczna - podziwiam fakt złożenia pracy doktorskiej z wykorzystaniem programu klasy MS Word i szczerze sugeruje na przyszłość wykorzystanie systemu Latex. O wiele łatwiej będzie zarządzać wszelkiego typu spisami i referencjami, a poza tym wszelkie wzory i równania będą zdecydowanie lepszej jakości.

Uwaga formalna - po zapoznaniu się z listą publikacji, na których Doktorant oparł swój manuskrypt, dochodzę do wniosku, że przedstawiony do recenzji doktorat mógłby w sposób zupełnie jednoznaczny być złożony w formie cyklu publikacji (dopuszczalnego przez ustawę, według której jest realizowane postępowanie Art. 13 p. 2) wraz z stosownym przewodnikiem. Wniosek ten motywuje obserwacją, iż Doktorant pełni wiodącą rolę w zdecydowanej większości publikacji związanych z pracą (7 publikacji, Doktorant jest pierwszym autorem w 5 z nich). Wszystkie publikacje są punktowane (zgodnie z aktualną listą Ministra Edukacji i Nauki) a trzy z nich posiadają współczynnik wpływu.

AJ

W manuskrypcie znalazłem szereg błędów literowych, niewielkich niedociągnięć językowych i typograficznych, które w żaden sposób nie wpłynęły na postrzeganie i wartość całej pracy.

Zgodnie z ustawą, według której jest realizowane postępowanie, tj. Ustawa z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki (Dz. U. 2003, nr 65, poz. 595, ze zm.)), Art.13. p. 1 i 2 Rozprawa doktorska (...) powinna stanowić oryginalne rozwiązanie problemu naukowego (...) oraz wykazywać ogólną wiedzę teoretyczną kandydata w danej dyscyplinie naukowej (...) oraz umiejętność samodzielnego prowadzenia pracy naukowej (...). Rozprawa doktorska może mieć formę maszynopisu książki, (...).. Stwierdzam, że przedstawiona do recenzji rozprawa spełnia przytoczone wymagania, gdyż Doktorant stosując metodę naukową przedstawił tło i motywację swoich badań, wskazał problemy istniejących rozwiązań, postawił tezę a następnie przedstawił rozumowanie oparte na eksperymentach dowodzące prawdziwości postawionej tezy. Niniejszym zwracam się więc do Rady Dyscypliny Naukowej Informatyka Techniczna i Telekomunikacja Politechniki Wrocławskiej o dopuszczenie mgr inż. Thanh Ngo Nguyena do dalszych etapów postępowania w sprawie nadania stopnia doktora nauk technicznych.