Prof. dr hab. Henryk Rybinski
Instytut Informatyki PW
Warszawa, Nowowiejska 17

Warszawa, 2023-09-08

Recenzja
rozprawy doktorskiej mgra inż Michała Kukowskiego
pt. "Indeksowanie baz danych na nowoczesnych typach pamięci".

1. Wstęp

Przedstawiona do recenzji rozprawa składa się z dwóch streszczeń – polskiego i angielskiego, ośmiu rozdziałów i bibliografii (233 pozycji, w tym 5 autorstwa bądź współautorstwa doktoranta). Objętość rozprawy - 185 stron. W swojej opinii przedstawię ogólną charakterystykę rozprawy, a następnie przedstawię swoje uwagi ogólne i szczegółowe.

2. Charakterystyka ogólna

2.1 Dziedzina

Obszarem badań zaprezentowanych w rozprawie jest dziedzina baz danych. Rozprawa jest poświęcona zagadnieniom indeksowania treści baz danych w nowoczesnych rodzajach pamięci masowych – w pojedynczych kościach typu flash, w dyskach SSD oraz pamięciach PCM.

Zagadnienia struktur danych na potrzeby indeksów baz danych towarzyszą rozwojowi baz danych od samego początku, przede wszystkim od momentu, gdy pojawiły się niesekwencyjne pamięci masowe. Wynikało to przede wszystkim z potrzeb budowania dużych systemów informacyjnych zapewniających efektywny dostęp do danych. Kierunki badań w tym zakresie są głównie wyznaczane przez dwa nurty:

1. z jednej strony rozwój technologiczny pamięci masowych,
2. z drugiej zaś intensywne badania naukowe w zakresie struktur danych.

Rozwój narzędzi związanych z metodami dostępu do danych w systemach baz danych jest w dziedzinie informatyki jednym z bardziej spektakularnych przypadków, pokazujących wzajemny wpływ badań naukowych i technologii na rozwój przemysłu bazodanowego.

Jednym z przełomów w tej dziedzinie było pojawienie się nowych rozwiązań technologicznych w zakresie pamięci masowych w latach 90-tych ubiegłego stulecia

Wraz z popularyzacją zastosowań półprzewodnikowych pamięci masowych pojawiły się pilne potrzeby zrewidowania rozwiązań znanych dla pamięci dyskowych i opracowania nowych metod. Wagę problematyki potwierdza literatura przedmiotu. Zapytanie *indexing methods in SSD* (a więc tylko dotyczące jednego typu pamięci rozważanych w rozprawie), do bazy Google Scholar zwraca blisko 31000 pozycji, z czego ponad 17000 publikacji baza GS wykazuje za okres od 2019 roku.

Opiniowana rozprawa wpisuje się w tematykę indeksowania baz danych w nowych strukturach pamięci masowych. Autor podejmuje próbę opracowania nowych metod indeksowania baz danych w pamięciach masowych oraz algorytmów utrzymywania struktur indeksowania.

Podsumowując tę część opinii, uważam, że tematyka rozważana przez doktoranta z praktycznego, jak też teoretycznego punktu widzenia jest ważna i jest godna rozprawy doktorskiej.

2.2 Konstrukcja rozprawy

Układ rozprawy jest logiczny i podporządkowany celom rozprawy. W konstrukcji pracy wyróżnić można następujące zasadnicze części:

1. motywacja badań, omówienie problemów badawczych poruszanych w pracy, przedstawienie celu rozprawy (rozdziały 1 - 3),
2. prezentacja warsztatu badawczego – omówienie zrealizowanej platformy testowej (Rozdział 4);

3. prezentacja opracowanych przez doktoranta nowych metod indeksowania oraz ocena ich efektywności (rozdziały 5 - 7);
4. Podsumowanie pracy (Rozdział 8)

Każdy z rozdziałów 5 - 7 jest poświęcony konkretnemu typowi pamięci:

1. Rozdział 5 wiąże się z prezentacją algorytmów dla pamięci typu *flash*;
2. W Rozdziale 6 zaprezentowano propozycje dla dysków SSD:
    a. przedstawiono propozycję nowej struktury indeksowania (drzewo FALSFM); dla tej struktury zaproponowano także algorytm dodawania zbioru rekordów.
    b. zaproponowano metodę indeksowania dla baz kolumnowych, wraz z algorytmem scalania atrybutów, charakteryzującym się lepszymi parametry niż metody istniejące. Ponadto przedstawiono analizę złożoności opracowanego algorytmu scalania;
3. Rozdział 7 jest poświęcony indeksowaniu w pamięciach PCM.

Prezentacji nowych algorytmów towarzyszą eksperymenty mające na celu porównanie proponowanych metod z metodami istniejącymi.

Rozprawę kończy podsumowanie oraz przedstawienie dalszych kierunków badań.

2.3 Charakter rozprawy

Jak wspomniałem, celem, jaki postawił sobie doktorant, jest analiza dotychczasowych algorytmów i struktur danych wykorzystywanych do indeksowania relacyjnych baz danych, a także opracowanie nowych metod indeksowania oraz implementacja zaproponowanych struktur danych. Dlatego też rozprawa ma przede wszystkim charakter praktyczny. Istotnym elementem rozprawy jest zbudowanie przez doktoranta platformy testowej. Platforma została zaimplementowana przez doktoranta z myślą o wprowadzeniu jednolitego środowiska do przeprowadzenia eksperymentów. Na podkreślenie zasługuje nowatorski symulator, który z jednej

strony służy do symulacji modeli pamięci, z drugiej zaś integruje algorytmy indeksowania w modelowanych modelach pamięci.

Opracowane rozwiązania są niewątpliwie istotne z punktu postawionych przez doktoranta celów, możliwości praktycznego wykorzystania symulatora są przekonywująco pokazane w rozdziałach 5-7.

Zawartość rozprawy dowodzi także solidnej wiedzy doktoranta w zakresie współczesnych technologii pamięci masowych.

Pozytywnie oceniam podjęte przez doktoranta próby analizy złożoności opracowanych algorytmów warsztat doktoranta, choć mam tu pewne zastrzeżenia do nie zawsze precyzyjnego definiowania używanych pojęć, np. nie znalazłem w pracy definicji pojęcia *amortyzowany czas* (pojawiło się w twierdzeniach 5.1 i 6.3).

3. Wkład autora

Wkład autora oceniam pozytywnie – obejmuje on szereg elementów związanych z metodami indeksowania danych w nowych pamięciach masowych. Wyróżnić tu można:

1. Zaimplementowanie zaawansowanego środowiska badawczego obejmującego symulację nowych rodzajów pamięci oraz dającego możliwość jednolitego prowadzenia analiz metod indeksowania dla różnych modeli pamięci.
2. Propozycja nowych metod indeksowania dla współczesnych pamięci masowych oraz przeprowadzenie eksperymentów dla zaproponowanych rozwiązań;
3. Teoretyczne oszacowanie efektywności zaproponowanych rozwiązań

Ad. 1

Zrealizowanie głównego celu, jaki postawił przed sobą autor, mianowicie przeanalizowanie istniejących rozwiązań dla nowych pamięci masowych oraz zbadanie efektywności zaproponowanych metod indeksowania wymagało stworzenia wiarygodnego warsztatu badawczego, tak aby możliwe było porównywanie poszczególnych metod w takim samym środowisku, a także wyeliminowanie

ewentualnych zniekształceń wyników, spowodowanych specyficznymi rozwiązaniami firmowymi (w tym m.in., optymalizatorami). Doktorant bardzo dobrze wywiązał się z tego zadania. Zaimplementowany został symulator pamięci masowych. Zrealizowany symulator integruje badane modele pamięci z algorytmami indeksowania. Autor wykorzystuje w badaniach benchmarki TCP-C i TCP-H.

Ad. 2

Zaproponowane nowe metody indeksowania dla pamięci Flash (FA-Tree oras LSM-Tree), SSD (indeks wierszowy FALSM-Tree, oraz indeks kolumnowy CF-Tree), a także PCM (BB+Tree oraz PCM Adaptive Merging). Dla wszystkich propozycji przeprowadzono eksperymenty, wykazując dużą przydatność opracowanych rozwiązań.

Ad. 3

Na uwagę zasługuje podjęta przez doktoranta próba teoretycznej analizy przedstawionych rozwiązań. W tym zakresie autor koncentruje się na dwóch rodzajach zagadnień:

1. Na analizie złożoności wybranych rozwiązań
2. Na analizie poprawności działania przyjętych algorytmów

W zakresie (1) w Rozdziale 5 oszacowano złożoność czasową algorytmu wstawiania elementu do drzewa (Twierdzenie 5.1), w Rozdziale 6 przeprowadzono analizę kosztów scalania kolumn (Twierdzenie 6.2) oraz oszacowano złożoność czasową algorytmu wstawiania rekordu do indeksu CFT (Twierdzenie 6.3).

W zakresie (2) udowodniono poprawność działania algorytmu wyszukiwania danych w przyjętej strukturze (Twierdzenie 6.1).

Na szczególną uwagę zasługują propozycje doktoranta dla pamięci SSD zawate w rozdziale 6, zarówno w zakresie baz danych wierszowych jak też kolumnowych.

4. Poprawność

Pozytywnie oceniam przyjętą przez autora metodologię badań. Autor dokonuje krytycznej analizy stanu badań w dziedzinie metod indeksowania w bazach danych oraz prezentuje technologię nowoczesnych pamięci masowych, w oparciu o tę analizę proponuje szereg własnych rozwiązań, a następnie przeprowadza badania eksperymentalne. Badania te stanowią istotny element rozprawy. Autor na danych benchmarkowych bada skonstruowane algorytmy i miary, porównując je z propozycjami z literatury. Uzyskane eksperymentalnie wyniki potwierdzają skuteczność proponowanych rozwiązań.

5. Inne uwagi

Uwagi ogólne

Niektóre z moich uwag poniżej mają charakter dyskusyjny, inne dotyczą redakcji rozprawy.

1. W bazach danych scenariusze dostępu do danych są bardzo różne. W swojej rozprawie doktorant skoncentrował się na bazach relacyjnych, Zastosowane w pracy benchmarki TPC są niewątpliwie najczęściej stosowane w tego typu badaniach, jako, że są bardzo zbliżone do typowych zastosowań (relacyjnych) baz transakcyjnych. Natomiast zabrakło mi w pracy choćby refleksji na temat hurtowni danych, czy też nierelacyjnych baz danych, takich jak np. bazy tekstowe, grafowe, czy XML.
2. Również pominięto aspekt dynamiki baz danych. W statycznych bazach wiedzy wymagania na szybkie wstawianie i usuwanie rekordów nie są tak istotne jak w bazach transakcyjnych. Dobrym miejscem dla takich rozważań mógłby być np. (rozbudowany?) punkt 6.3.2. W szczególności pojawia się pytanie, czy nie należałoby zaproponować rozwiązania dla różnych charakterystyk operacji na zbiorze, inny dla hurtowni (dwa rodzaje aktywności: (1) masowe dodawanie danych i (2) ograniczenie strumienia zadań do wyszukiwania), inny dla baz słownikowych (niewielkie zmiany), inne dla baz transakcyjnych. Pojawia się tu konkretne pytanie - czy autor się zastanawiał nad tym, aby dynamicznie określać wartość T dla określonej charakterystyki strumienia zadań.

3. Pewnym mankamentem jest też brak formalnego opisu B i B+ drzew w części wprowadzającej rozprawy, brak pewnych formalnych definicji (np. moja uwaga dot. amortyzowanego czasu), a także brak wykazu oznaczeń.
4. Brak w rozprawie wykazu stosowanej notacji oraz wykaz rysunków i tabel; wprowadzenie wykazów ułatwiłoby poruszanie się po tekście

Ważniejsze uwagi szczegółowe

Mam wątpliwości, co do p. 2.3, gdzie autor łączy problem efektywności budowania indeksu z problemem optymalnego wyboru indeksów (ang. *Index Selection Problem*). Sądzę, że ten temat zbyt odbiega od głównego nurtu rozprawy, jako, że jest to sam w sobie problem i można go rozważać abstrahując od metod indeksowania (por. np. rozprawa doktorska P. Kołaczkowskiego z roku 2011:

> „Autonomiczny dobór indeksów w systemach relacyjnych baz danych przy pomocy ewolucyjnego przekształcania planów wykonania zapytań"

Myślę, że w istniejącej formie punkt 2.3 raczej wprowadza zamieszanie, niż coś wyjaśnia.

W Tabeli 3.1 sugerowałbym jednolitą miarę we wszystkich wierszach

Tytuł rozdziału 3.1 powinien odwoływać się do pamięci masowej, nie do dysków (SSD jest nazywane dyskiem mocno z rozpędu, także w literaturze przedmiotu)

W Tabeli 3.3 w kolumnie 1 autor pisze o poborze energii, a w kolumnie 3 pojawiają się jednostki prądu – zbyt duży przeskok myślowy;

Str 39 – w punkcie 2 autor pisze o pielęgnacji:

„...druga faza zaczyna się, gdy czas jest mniejszy niż...", rozumiem, że chodzi o czas dostępu;

Str 39 uwaga terminologiczna – proponowałbym zastąpić tłumaczenie terminu *harvesting* jako „zbiory" terminem „żniwa" (w polskiej terminologii „zbiory" kojarzą się z teorią zbiorów)

W pracy brakuje niektórych definicji podstawowych. Na str. 40 pojawia się termin „selektywność", ale brak definicji tego terminu (Uwagi ogólne – uwaga 3)

Podobnie z pojęciem *czas amortyzowany* w twierdzeniach 5.1 i 6.3

Zbyt powierzchownie został potraktowany przez doktoranta rozdział 8. Uważam, że rozprawa zasługuje na głębsze podsumowanie.

W trakcie czytania rozprawy miałem też szereg uwag dotyczących stylu i interpunkcji. W szczególności, gdyby doktorant zamierzał opublikować monografię należałoby wykonać porządną adjustację tekstu. Niektóre moje uwagi w tym zakresie znajdują się w pliku PDF (przekazane autorowi do wykorzystania przy ewentualnym przygotowaniu monografii).

Powyższe moje uwagi nie umniejszają pozytywnej oceny rozprawy.

Podsumowanie

Praca ma oryginalny charakter. Wkład autora jest znaczący. Jest to bardzo ciekawa propozycja w zakresie nowych metod indeksowania we współczesnych pamięciach masowych. Dlatego uważam, że opiniowana praca spełnia wymagania zawarte w obowiązujących przepisach dotyczących rozpraw doktorskich, wnoszę zatem o dopuszczenie mgra inż. Michała Kukowskiego do publicznej obrony.