Politechnika Warszawska
Wydział Elektroniki i Technik Informacyjnych

Warszawa, 26 lutego 2018 r.

D z i e k a n a t

Uprzejmie informuję, że na Wydziale Elektroniki i Technik Informacyjnych Politechniki Warszawskiej odbędzie się w dniu 13 marca 2018 r. publiczna obrona rozprawy doktorskiej

**mgr inż. Marka Łatuszko**

**temat:** :"Methods for Solving the View Selection Problem in Data Cubes"
„Metody rozwiązywania zadania wyboru widoków w kostkach danych"

**promotor – prof. dr hab. inż. Radosław Pytlak z Wydziału Matematyki i Nauk Informacyjnych Politechniki Warszawskiej**

**recenzenci:**
**prof. dr hab. inż. Kazimierz Subieta z Polsko-Japońskiej Akademii Technik Komputerowych**
**prof. dr hab. inż. Zbyszko Królikowski z Politechniki Poznańskiej**

Obrona odbędzie się w dniu 13 marca 2018 r. w sali 116 na Wydziale Elektroniki i Technik Informacyjnych – Gmach im. Janusza Groszkowskiego, Warszawa, ul. Nowowiejska 15/19; początek godz. 12.00 .

Po adresem: **www.elka.pw.edu.pl/Wydzial/Rada-Wydzialu/Harmonogram-obron-doktorskich-streszczenia-i-recenzje** zapewniony jest na stronie Wydziału dostęp do tekstów streszczenia rozprawy i recenzji, jak również do tekstu rozprawy umieszczonej w Bazie Wiedzy Politechniki Warszawskiej.

Dziekan

prof. dr hab. inż. Krzysztof Zaremba

Mgr inż. Marek Łatuszko

Promotor – prof. dr hab. inż. Radosław Pytlak

Tytuł rozprawy doktorskiej” Methods for Solving the View Selection Problem in Data Cubes”

„Metody rozwiązywania zadania wyboru widoków w kostkach danych".

STRESZCZENIE. Jedną z najbardziej znaczących i powszechnie stosowanych metod przyspieszania realizacji zapytań użytkowników w wielowymiarowych bazach danych jest materializacja widoków. Problem wyboru do materializacji właściwej części struktury danych przy ograniczonych zasobach znany jest jako problem wyboru widoków. W niniejszej rozprawie doktorskiej badane są dwa zadania wyboru widoków: minimalizacja średniego czasu wykonywania zapytań przy ograniczonej wielkości pamięci dostępnej do przechowywania zmaterializowanych widoków oraz minimalizacja wielkości pamięci wymaganej do przechowywania zmaterializowanych widoków przy spełnieniu warunku nieprzekroczenia średniego czasu wykonywania zapytań. Pierwsze zadanie zostało określone z perspektywy administratora bazy danych, podczas gdy drugie jest zdefiniowane z punktu widzenia twórcy oprogramowania. Oba zadania zostały sformułowane jako zadania optymalizacji całkowitoliczbowej, co pozwoliło na zastosowanie algorytmów optymalizacji całkowitoliczbowej oraz oprogramowania rozwiązującego tego typu zadania. W ramach rozprawy doktorskiej zostały zbadane różne metody heurystyczne, bazujące na algorytmie zachłannym, wykazane twierdzenia dotyczące tych metod oraz zaproponowane ich modyfikacje, które wpływają zarówno na skrócenie czasu znajdowania rozwiązań, jak też na zmniejszenie wartości funkcji celu. Dodatkowo, zostały przeprowadzone testy numeryczne porównujące (pod względem czasu działania i kosztu znalezionych rozwiązań) metody heurystyczne oraz powszechnie stosowane oprogramowanie służące do rozwiązywania zadań całkowitoliczbowych. Tym co wyróżnia to porównanie jest jego kompleksowość, która została osiągnięta dzięki zastosowaniu profili wydajności. Zostały także zaproponowane dwie metody redukcji złożoności obliczeniowej, które znacząco przyspieszają działanie algorytmów heurystycznych i dokładnych, bez zwiększenia wartości funkcji celu. Zaprezentowane eksperymenty zostały przeprowadzone na rozbudowanym zbiorze zadań testowych, ze szczególnym uwzględnieniem dużych zadań, rzadko rozważanych w dotychczasowych doświadczeniach.

Poznań, 27.10.2017

prof. dr hab. inż. Zbyszko Królikowski
Politechnika Poznańska
Wydział Informatyki
Instytut Informatyki
ul. Piotrowo 2
60-965 Poznań
e-mail: Zbyszko.Krolikowski@cs.put.poznan.pl

# Recenzja rozprawy doktorskiej

## mgr. inż. Marka Łatuszko

## pt. *Methods for Solving the View Selection Problem in Data Cubes*

## 1. Tematyka i zarys problemu

Recenzowana rozprawa doktorska mgr. inż. Marka Łatuszko jest poświęcona problematyce doboru najlepszego zbioru perspektyw zmaterializowanych dla optymalizacji zadanego obciążenia. Obciążenie rozumie się tu jako zbiór zapytań analitycznych. Wybór tego zbioru podlega pewnym ograniczeniom, co stanowi istotę problemu. Rozprawa dotyczy znanego problemu, trudnego obliczeniowo, którego rozwiązaniem zajmują się główne ośrodki naukowe na świecie od początku lat 90-tych.

Perspektywa zmaterializowana (ang. *materialized view*) jest strukturą danych trwale przechowującą wynik pewnego zapytania $q_i$. Celem jej stosowania jest skrócenie czasu wykonania złożonego, a tym samym czasochłonnego zapytania $q_i$, które zamiast na oryginalnych relacjach (tabelach) może zostać wykonane szybciej na perspektywie zmaterializowanej lub na zbiorze takich perspektyw. Ten proces jest nazywany przepisywaniem zapytań (ang. *query rewriting*) i jest stosowany w praktyce w komercyjnych systemach zarządzania bazą danych.

Perspektywy zmaterializowane i przepisywanie zapytań są jednymi z podstawowych mechanizmów optymalizacji zapytań w hurtowniach danych.

W ogólności, główna trudność koncepcji wykorzystujących perspektywy zmaterializowane polega na znalezieniu takiego ich zbioru $M$, który z jednej strony, będzie mógł być wykorzystany do optymalizacji jak największego zbioru najbardziej złożonych obliczeniowo zapytań, a z drugiej strony, zaproponowany zbiór $M$ będzie mógł zostać odświeżony w zadanym oknie czasowym. Ponadto,

Poznań, 27.10.2017

prof. dr hab. inż. Zbyszko Królikowski
Politechnika Poznańska
Wydział Informatyki
Instytut Informatyki
ul. Piotrowo 2
60-965 Poznań
e-mail: Zbyszko.Krolikowski@cs.put.poznan.pl

# Recenzja rozprawy doktorskiej

## mgr. inż. Marka Łatuszko

## pt. *Methods for Solving the View Selection Problem in Data Cubes*

## 1. Tematyka i zarys problemu

Recenzowana rozprawa doktorska mgr. inż. Marka Łatuszko jest poświęcona problematyce doboru najlepszego zbioru perspektyw zmaterializowanych dla optymalizacji zadanego obciążenia. Obciążenie rozumie się tu jako zbiór zapytań analitycznych. Wybór tego zbioru podlega pewnym ograniczeniom, co stanowi istotę problemu. Rozprawa dotyczy znanego problemu, trudnego obliczeniowo, którego rozwiązaniem zajmują się główne ośrodki naukowe na świecie od początku lat 90-tych.

Perspektywa zmaterializowana (ang. *materialized view*) jest strukturą danych trwale przechowującą wynik pewnego zapytania $q_i$. Celem jej stosowania jest skrócenie czasu wykonania złożonego, a tym samym czasochłonnego zapytania $q_i$, które zamiast na oryginalnych relacjach (tabelach) może zostać wykonane szybciej na perspektywie zmaterializowanej lub na zbiorze takich perspektyw. Ten proces jest nazywany przepisywaniem zapytań (ang. *query rewriting*) i jest stosowany w praktyce w komercyjnych systemach zarządzania bazą danych.

Perspektywy zmaterializowane i przepisywanie zapytań są jednymi z podstawowych mechanizmów optymalizacji zapytań w hurtowniach danych.

W ogólności, główna trudność koncepcji wykorzystujących perspektywy zmaterializowane polega na znalezieniu takiego ich zbioru $M$, który z jednej strony, będzie mógł być wykorzystany do optymalizacji jak największego zbioru najbardziej złożonych obliczeniowo zapytań, a z drugiej strony, zaproponowany zbiór $M$ będzie mógł zostać odświeżony w zadanym oknie czasowym. Ponadto,

często ogranicza się przestrzeń pamięci dyskowej dostępnej dla składowania zbioru $M$. Te trzy ograniczenia powodują, że problem wyznaczenia optymalnego zbioru $M$ jest trudny obliczeniowo i w praktyce do jego rozwiązania stosuje się algorytmy przybliżone.

W tym kontekście mgr. inż. Marek Łatuszko podjął się **rozwiązania nietrywialnego problemu badawczego**.

## 2. Struktura rozprawy

Recenzowana rozprawa doktorska składa się z 8 rozdziałów, 2 załączników i bibliografii, o łącznej objętości 124 stron (rozprawa - 80 stron, bibliografia - 6 stron, załączniki - 24 strony, strony początkowe i spisy - 14 stron). Rozdział 1 zawiera wprowadzenie do problematyki, motywację do podjęcia problemu oraz cel i zakres rozprawy. Rozdziały 2 i 3 wprowadzają podstawowe koncepcje wykorzystywane w rozprawie. Rozdział 4 zawiera sformułowanie problemu rozwiązywanego w dalszej części rozprawy, tj. znalezienie zbioru perspektyw zmaterializowanych spełniający przyjęte założenia. Rozdziały 5, 6 i 7 stanowią autorski wkład Doktoranta w problematykę wyboru optymalnego zbioru perspektyw zmaterializowanych. Rozdział 5 omawia algorytm znajdujący zbiór perspektyw zmaterializowanych $M$ o najmniejszym rozmiarze, który zminimalizuje średni czas wykonania zapytania przy ograniczeniu rozmiaru przestrzeni dyskowej przeznaczonej na składowanie perspektyw ze zbioru $M$. Rozdział 6 omawia algorytm znajdujący zbiór perspektyw zmaterializowanych $M$ o minimalnym rozmiarze, przy ograniczeniu średniego czasu wykonania zapytania na perspektywach ze zbioru $M$. Rozdział 7 weryfikuje zależność korzyści (np. czas wykonania) ze zmaterializowania perspektyw w zależności od sumarycznego rozmiaru tych perspektyw. Rozdział 8 podsumowuje rozprawę. W załącznikach zostały umieszczone: (1) opis kostki danych, (2) rozszerzone wyniki eksperymentalne, nieujęte w rozdziale 7.

## 3. Wyniki rozprawy

Przedmiot recenzowanej rozprawy zalicza się do ważnego nurtu badań na świecie w dziedzinie wyboru zbioru perspektyw zmaterializowanych optymalizujących zadany zbiór zapytań, przy pewnych ograniczeniach przestrzeni dyskowej na składowanie tych perspektyw. W literaturze naukowej problem ten nazywa się *view selection* problem.

Do głównych wyników rozprawy zaliczam:

- Zaproponowane w Rozdziale 5 usprawnienia istniejących algorytmów zachłannych znajdujące zbiór perspektyw zmaterializowanych $M$, przy ograniczeniu przestrzeni dyskowej. Usprawnienia te dotyczą redukcji przestrzeni przeszukiwania rozwiązań i ograniczenia

wykorzystania pamięci. Doktorant przeanalizował teoretycznie złożoność obliczeniową tych algorytmów, zaimplementował je i porównał eksperymentalnie z *solverem* programowania całkowito-liczbowego. Otrzymane wyniki potwierdziły zasadność zastosowania usprawnień zaproponowanych w rozprawie.

- Zaproponowany w Rozdziale 6 algorytm wprowadzający usprawnienie do algorytmu minimalizującego rozmiar zbioru $M$, przy ograniczeniu na średni czas wykonania zapytań z zadanego zbioru obciążenia $Q$. Podobnie jak w poprzednim rozdziale, Doktorant przeanalizował teoretycznie zachowanie się algorytmu i porównał go z rozwiązaniami alternatywnymi i podobnie jak poprzednio z *solverem* programowania całkowito-liczbowego. Otrzymane wyniki eksperymentalne potwierdziły zasadność zaproponowanego rozwiązania.
- Eksperymentalne zweryfikowanie znanej z literatury naukowej zależności kosztu wykonania zapytań z zadanego zbioru obciążenia $Q$ w zależności od wolumenu zmaterializowanych danych w zbiorze $M$.

## 4. Uwagi

### Rozwiązywany problem

W literaturze światowej problem, którym zajął się Doktorant jest badany od lat 90-tych. Zaproponowano wiele algorytmów, które głównie zaliczają się do klasy zachłannych (ang. *gridy*). Algorytmy te znajdują (sub)-optymalne zbiory perspektyw zmaterializowanych $M$, optymalizujące wykonanie zadanego zbioru zapytań $Q$. Autorzy tych algorytmów przyjmują dwa założenia upraszczające. Po pierwsze, zakładają że prawdopodobieństwo wystąpienia każdego zapytania w zbiorze $Q$ jest identyczne. Po drugie, w modelu kosztów nie uwzględniają czasów odświeżania perspektyw ze zbioru $M$, a jedynie ograniczają rozmiar przestrzeni dyskowej na składowanie $M$. Tą samą drogą podążył Doktorant proponując usprawnienia algorytmów zachłannych.

Zdaniem autora niniejszej opinii, bardziej wartościowym byłoby podjęcie próby rozwiązania wspomnianego *view selection problem* budując model, w którym: (1) prawdopodobieństwo wystąpienia każdego zapytania ze zbioru $Q$ miałoby niezależny rozkład, (2) model kosztów uwzględniałby koszty odświeżania perspektyw ze zbioru $M$. W recenzowanej rozprawie brakuje dyskusji na ten temat, nawet jeśli nie podjęto próby rozwiązania tego problemu.

W tym kontekście nasuwa się pytanie, czy wykorzystane algorytmów zachłannych omówionych w niniejszej rozprawie nadal miałoby sens przy tak rozszerzonym *view selection problem*.

## Stan wiedzy

W rozprawie Doktorant odwołuje się do wielu prac z zakresu algorytmów doboru perspektyw materializowanych. Następujące prace nie zostały jednak ujęte w stanie wiedzy:

- de Souza M. F., Sampaio M. C.: Efficient materialization and use of views in data warehouses. SIGMOD Record, (28):1, 1999
- Theodoratos D., Xu W.: Constructing search spaces for materialized view selection. DOLAP, 2004
- Xu W., Theodoratos D., Zuzarte C.: Computing closest common subexpressions for view selection problems. DOLAP, 2006
- Boukorca A., Bellatreche L., Senouci S-A.B., Faget Z.: Coupling Materialized View Selection to Multi Query Optimization: Hyper Graph Approach. IJDWM 11(2), 2015

Ponadto, problematyka związana z budowaniem *data-cube* dotyczy wykorzystania agregatów dla optymalizacji zapytań z operatorami GROUP BY CUBE i GROUP BY ROLAP. W tym zakresie, w literaturze światowej zaproponowano kilka algorytmów (np. *PipeSort, PipeHash, Overlap, Partitioned Cube, Memory Cube, Bottom-Up Computation*). Algorytmy te mogłyby zostać rozważone do rozwiązania problemów poruszanych w rozprawie. Referencje do najważniejszych ze wspomnianych prac są następujące:

- Agarwal S., Agrawal R., Deshpande M. P., Gupta A., Naughton F. J., Ramakrishnan R., Sarawagi S.: On the Computation of Multidimensional Aggregates. VLDB, 1996
- Ross K. A., Srivastava D.: Fast Computation of Sparse Datacubes. VLDB, 1997
- Beyer K, Ramakrishnan R.: Bottom-Up Computation of Sparse and Iceberg CUBEs. SIGMOD, 1999

## Edycja i skład tekstu

Pod względem edytorskim rozprawa została przygotowana bardzo dobrze - ma tylko kilka następujących uchybień.

- Bibliografia: formatowanie pozycji bibliograficznych jest niespójne. Przykładowo, [2] i [3] dotyczą artykułów publikowanych na dwóch różnych konferencjach VLDB, ale te pozycje bibliograficzne mają inne formaty. Podobna uwaga dotyczy [17], [19], [23].
- W pozycjach dotyczących wielu konferencji VLDB pojawia się tekst 'San Francisco, CA, USA', który w tych przypadkach nie oznacza miejsca odbywania się konferencji, natomiast w [13] - 'Shanghai, China' oznacza miejsce odbywania się konferencji.

- Niektóre nazwy czasopism, por. np. [28], [38] są zapisane skrótami, natomiast inne - pełnymi nazwami, por. np. [37], [51].
- Na stronie 44, Tabela 5.1 powinna się znaleźć za tekstem odnoszącym się do niej.
- Zgodnie ze standardami składu tekstu, podrozdziały wstawiamy tylko, gdy jest ich więcej niż 1. Nie jest to przypadek 3.1, który w rozdziale 3 występuje jako jedyny.

## 5. Ocena końcowa

Podsumowując, uważam, że cel rozprawy został osiągnięty. Mgr inż. Marek Łatuszko wykazał, że zastosowanie zaproponowanych algorytmów, przy założonych ograniczeniach, może poprawić ogólną jakość (w zależności od podejścia może to być minimalizacja rozmiaru) znalezionego zbioru perspektyw zmaterializowanych lub skrócić czas znajdowania tego zbioru. Doktorant wykazał się dobrą znajomością stanu badań w tematyce rozprawy, biegłą znajomością zaawansowanego aparatu matematycznego, analitycznych metod szacowania złożoności obliczeniowej, umiejętnością rozwiązywania problemów badawczych i wiedzą praktyczną z zakresu implementowania algorytmów.

Doktorant osiągnął oryginalne wyniki naukowe, których jakość została potwierdzona w publikacjach w czasopiśmie naukowym European Journal of Operational Research (40 pkt.) oraz Journal of Intelligent Information Systems (20 pkt.).

W tym kontekście, **uważam, że recenzowana rozprawa doktorska spełnia z wyraźnym nadmiarem wymagania stawiane rozprawom doktorskim przez obowiązującą ustawę, wobec czego wnoszę o dopuszczenie jej do publicznej obrony**.

**tytuł, stopień, imię i nazwisko**
Prof. dr hab. inż. Kazimierz Subieta

**data**
20 listopada 2017 r.

**miejsce pracy**
Polsko Japońska Akademia Technik Komputerowych
ul. Koszykowa 86, 02-008 Warszawa

***KWESTIONARIUSZ- RECENZJA ROZPRAWY DOKTORSKIEJ DLA RADY WYDZIAŁU ELEKTRONIKI I TECHNIK INFORMACYJNYCH POLITECHNIKI WARSZAWSKIEJ***

**Tytuł rozprawy:** Methods for Solving the View Selection Problem in Data Cubes
(*Metody rozwiązywania problemu selekcji widoków w kostkach danych*)

**Autor rozprawy:** mgr inż. Marek Łatuszko

1. **Jakie zagadnienie naukowe jest rozpatrywane w pracy (teza rozprawy) i czy zostało ono jasno sformułowane przez autora? Jaki charakter ma rozprawa (teoretyczny, doświadczalny, inny)?**

W szerokim znaczeniu, praca doktorska mgr inż. Marka Łatuszko ma na celu usprawnienie efektywności procesów decyzyjnych opartych na eksploracji danych z zastosowaniem kostki danych.

W węższym znaczeniu, praca ma na celu opracowanie algorytmów, których celem jest minimalizacja czasu odpowiedzi na zapytania użytkownika skierowane do kostki danych lub minimalizacja pamięci niezbędnej do realizacji takich zapytań. Oba te cele zostały ujęte w modelach matematycznych (nieco różnych w zależności od celu), zaś przedstawione metody opracowane przez autora rozprawy mają gwarantować osiągnięcie tych celów.

Zagadnienie naukowe zostało więc sformułowane w sposób jasny, jakkolwiek wątpliwości (o których dalej) mogą dotyczyć zbyt wąskiego sformułowania tych modeli i ich usytuowania w całości procesów decyzyjnych bazujących na kostkach danych.

Praca ma zdecydowanie charakter teoretyczny i nie dokumentuje jakichkolwiek prób zastosowania jej wyników w rzeczywistych systemach informatycznych wspomagających podejmowanie decyzji i/lub wniosków wynikających z takich prób. Nie ma w niej także wzmianki o tym, że autor pracy miał kiedykolwiek do czynienia z rzeczywistymi zastosowaniami kostki danych w praktyce i zdobył na tym terenie istotne doświadczenia. Na usprawiedliwienie autora dodam, że zdobycie takich doświadczeń oraz przeprowadzenie takich prób w warunkach akademickich jest praktycznie niewykonalne. To jednak rodzi zasadnicze pytanie, już nie do doktoranta, ale raczej do Rady Wydziału: czy na uczelni technicznej

powinny być podejmowane badania, o których z góry wiadomo, że temat powstał wyłącznie na bazie rozeznania literaturowego zaś rezultaty będą nieweryfikowalne w praktyce?

2. **Czy w rozprawie przeprowadzono w sposób właściwy analizę źródeł (w tym literatury światowej, stanu wiedzy i zastosowań w przemyśle) świadczącej o dostatecznej wiedzy autora? Czy wnioski z przeglądu źródeł sformułowano w sposób jasny i przekonywujący?**

Literatura na temat budowy, własności i zastosowań kostek danych, a szerzej na temat eksploracji danych (*data mining*), jest bardzo bogata i obejmuje setki lub tysiące pozycji. Autor zacytował 77 pozycji, co wydaje się rozsądną liczbą. Nie potrafię się wypowiedzieć, czy ten wybór jest reprezentatywny i optymalny. Tak czy inaczej, czytelnik, który zajmie się tym lub podobnym tematem, będzie polegać raczej na powszechnie dostępnych bibliografiach (Google, DBLP, itd.), a nie wyłącznie na wyborze dokonanym przez autora. Ten wybór posłużył do skonstruowania krótkiego (2++ strony) podrozdziału rozprawy „*Related work*", w którym omówiono niektóre zagadnienia poruszane w literaturze oraz podano odpowiednie odsyłacze. Omówienie jest lakoniczne i sprowadza się w zasadzie do pojedynczych zdań charakteryzujących dany temat. Autor nie podejmuje prób krytycznej oceny stanu badań i wiedzy w poszczególnych tematach, jest to raczej krótka charakterystyka ich treści. Tymczasem moje doświadczenia z tym tematem pokazują, że większość tej literatury należy do gatunku określanego eufemistycznie jako „*pure theoretical art*" i nie zmierza do jakichkolwiek zastosowań. W recenzowanej pracy przemysłowe zastosowania tych technologii są dyskutowane dość pobieżnie. Nie dopatrzyłem się też istotnych wniosków z analizy literatury, poza oczywistym wnioskiem, że jest pożądane, aby operacje na kostce danych przebiegały w sposób bardziej sprawny.

3. **Czy autor rozwiązał postawione zagadnienia, czy użył właściwej do tego metody i czy przyjęte założenia są uzasadnione?**

Autor na wiele sposobów ogranicza rozpatrywane przez niego środowisko wspomagania decyzji z zastosowaniem kostki danych. Jest to naturalne w badaniach naukowych, gdyż często rzeczywisty problem jest zbyt złożony, aby rozpatrywać go w całości. Konieczne są abstrakcje, które pozwolą na ograniczenie przestrzeni aspektów problemu i pozwolą na wyciągnięcie daleko idących i nietrywialnych wniosków. Ten oczywisty proces tworzenia środowiska badań naukowych kryje w sobie niebezpieczeństwo zerwania związków z rzeczywistością. To zdarzyło się w wielu działach informatyki, np. w inżynierii oprogramowania, którą próbowano traktować np. przy pomocy logiki matematycznej lub semantyki denotacyjnej ze skutkiem nie mającym nawet najdrobniejszego znaczenia dla rzeczywistego postępu w tej dziedzinie. Przykłady można mnożyć w różnorodnych dziedzinach informatycznych, gdzie akademickie badania naukowe całkowicie rozminęły się z rzeczywistymi problemami i potencjalnymi zastosowaniami.

Ponieważ autor nie dokumentuje w tej pracy rzeczywistych zastosowań opracowanych metod i algorytmów nie da się jednoznacznie ustalić, że poczynione przez niego założenia ograniczające problem są prawidłowe i nie prowadzą do bezpłodnej scholastyki. Jako specjalista z zakresu inżynierii oprogramowania i baz danych mam jednak pewne niepokojące wrażenia. Pierwsze z nich, absolutnie zasadnicze, polega na tym, że w pracy nie pojawia się

użytkownik wraz z jego potrzebami i preferencjami. Optymalizacje rozpatrywane w pracy dotyczą wyboru widoków (*views*) na podstawie kosztu ewaluacji zapytania rozpatrywanego jako abstrakt, co całkowicie zaniedbuje fakt, że widoki i zapytania są elementami interfejsu użytkownika i on jest jedynym suwerenem przesądzającym o wyborze ich postaci. Z tego punktu widzenia optymalizacje rozpatrywane w pracy można uznać z założenia za kontrowersyjny pomysł, lub wręcz za pobawione jakiegokolwiek sensu. Dla użytkownika nie jest wszystko jedno, czy widok, z którym ma pracować, zawiera np. informacje o klientach z pominięciem dostawców, czy może odwrotnie. Można sobie wyobrazić, że optymalizacje proponowane przez autora mają tylko wspomóc użytkownika w wyborze widoków, ale ta okoliczność nie występuje w pracy i jest raczej moją ekstrapolacją intencji autora. Wybór widoków silnie zależy od charakterystyki zapytań do kostki danych (tak jest w bazach danych), ale ten fakt również w pracy nie jest rozpatrywany.

Wobec tego próbuję zrozumieć tę pracę od innej strony. Jak autor sugeruje, widoki są w tej pracy wyłącznie strukturami pomocniczymi służącymi jako wspomaganie ewaluacji zapytań, które zawsze są kierowane do podstawowej kostki danych. Użytkownik z takimi widokami nie ma nic do czynienia, jest to wewnętrzna sprawa mechanizmu ewaluacji zapytań. Takie założenie rodzi jednak kolejne wątpliwości. Drobna wątpliwość dotyczy terminologii: to nie są widoki tak jak są one rozumiane powszechnie. Widoki, jak nazwa wskazuje, należą do interfejsu użytkownika, co w tym przypadku nie ma miejsca. Nie wykluczone, że w środowisku osób zajmujących się eksploracją danych wykształciło się takie (nad)użycie tego terminu. Nie jest to jednak właściwe i może być mylące dla wielu czytelników tej pracy.

Pomijając terminologię, sprawa takiego traktowania tego rodzaju struktur pomocniczych rodzi znowuż dość zasadnicze pytania. Optymalizacja zapytań (np. w SQL) kierowanych do bazy danych wykształciła ogromną różnorodność pomocniczych struktur danych i innych mechanizmów mających na celu zmniejszenie czasów odpowiedzi na zapytania. Dotyczy to w szczególności różnorodnych indeksów, *access support relations*, map bitowych, zapamiętanych wyników zapytań (*cached queries*), metod opartych na kodowaniu mieszającym (*hash coding*), na sortowaniu, itd. Powstaje pytanie, dlaczego autor w swoich badaniach ograniczył się do jednego rodzaju takich struktur powstających poprzez rzutowanie (ze specyficzną agregacją) n-wymiarowej kostki danych na mniejszą liczbę wymiarów? Z czego to ograniczenie wynika i czy rzeczywiście takie ograniczenie jest na tyle powszechne, że warto mu poświęcać specjalne badania?

Nie dysponuję danymi statystycznymi w tym zakresie, być może tak jest, ale to może być sprawdzone wyłącznie poprzez eksperymenty w rzeczywistym środowisku systemu podejmowania decyzji. Jak wspomniałem, praca nie dokumentuje tego rodzaju eksperymentów. Zatem, jeżeli całość jest oparta na spekulacjach *a priori*, to można byłoby również rozpatrywać inne operatory w języku zapytań, nie tylko wspomniany operator rzutowania. W szczególności, takim operatorem może być operator selekcji, który w zastosowaniach kostek danych jest równie ważny, a nawet niekiedy ważniejszy od operatora rzutowania. Np. jeżeli ktoś z działu marketingu zajmuje się sprzedażą butów w miesiącach zimowych, to nie jest dla niego interesująca sprzedaż masła w lipcu. Dla tego przypadku potrzebny jest indeks towarów oraz odpowiedni widok powstający z zastosowaniem operatora selekcji. Niestety, ta wątpliwość doprowadzi większość czytelników do wniosku, że to ograniczenie zostało podyktowane przyjętym modelem matematycznym, który zdominował

myślenie o problemie. Rozszerzenie problemu prowadzi do nieadekwatności tego modelu matematycznego lub wręcz (najczęściej) do niemożliwości zbudowania takiego modelu. Tego rodzaju ogranicznia są przyczyną bardzo niskiej skuteczności badań naukowych oraz upadku autorytetu badań akademickich w środowiskach przemysłowych w ważnych działach informatyki (np. w inżynierii oprogramowania, bazach danych, językach programowania, technologii *workflow*, *middleware*, itd.). Ponieważ część mojego życia zawodowego spędziłem w środowiskach przemysłowych, znam ich mocno lekceważący stosunek do badań akademickich. Przykładowo, przemysłowe konsorcjum standaryzacyjne OMG (odpowiedzialne za standardy CORBA, UML i inne), w którego pracach uczestniczyłem przez kilka lat, dopuszcza, aby przedstawiciele środowisk akademickich brali udział w pracach poszczególnych komitetów, ale nie daje im praw do głosowania nad poszczególnymi cechami danego standardu. Takie prawo mają wyłącznie przedstawiciele firm komercyjnych. Nie ma wątpliwości, jest to bezpośrednia konsekwencja niskiej oceny badań akademickich z punktu widzenia zastosowań praktycznych, połączonej z niewiarą w akademickie autorytety.

Aby nieco złagodzić wymowę ostatniego akapitu i nie wpadać w idealizującą przesadę można dodać, że ścieżki odkryć naukowych nie są proste i wymagają wielu prób, udanych i nieudanych. Z tego punktu widzenia tę rozprawę doktorską można traktować jako ćwiczenie w zakresie zawężonego problemu niezbędne do tego, aby zająć się szerszym i bardziej rzeczywistym problemem.

4. **Na czym polega oryginalność rozprawy, co stanowi samodzielny i oryginalny dorobek autora, jaka jest pozycja rozprawy w stosunku do stanu wiedzy czy poziomu techniki reprezentowanych przez literaturę światową?**

Autor rozprawy zaproponował oryginalne metody wyboru widoków bazujących na kostce danych, które optymalizują czas odpowiedzi na zapytania lub pamięć niezbędną do realizacji zapytań. Niestety, nie jestem specjalistą z zakresu matematycznych metod optymalizacyjnych, więc nie będę wypowiadać się na temat jakości wywodów matematycznych w tej rozprawie. Pozostaje mi wierzyć, że w tym zakresie autor wykonał solidną pracę. Całość sprawia bardzo dobre wrażenie. Nie jestem w stanie również sprawdzić, czy opracowane przez autora algorytmy są rzeczywiście lepsze od tych znanych z literatury. Nie dopatrzyłem się w pracy fragmentów porównujących proponowane algorytmy z propozycjami innych autorów. Nie traktuję tego jako wadę pracy: takie porównanie jest po pierwsze bardzo trudne ze względu na złożoną implementację i testy, po drugi może być niewykonalne ze względu na różne aspekty i różne cele adresowane przez różne algorytmy. Pobieżny przegląd pozostawia wrażenie, że generalnie w zakresie optymalnej selekcji widoków literatura nie jest obfita, więc nowe rezultaty są bardzo pożądane. Zgodnie z moją wiedzą, podjęcie tego tematu jest już samo w sobie oryginalne.

5. **Czy autor wykazał umiejętność poprawnego i przekonywującego przedstawienia uzyskanych przez siebie wyników (zwięzłość, jasność poprawność redakcyjna rozprawy)?**

Praca na ogół jest jasna i poprawnie zredagowana. Części pracy przedstawiające model matematyczny mogą powodować trudności u czytelników mniej przywiązanych do badań o charakterze matematycznym, ale jest to sprawa nieuchronna, wynikająca z matematycznej

specyfiki tej pracy. Dla nie-matematyków, szczególnie dla osób ze środowisk przemysłowych, przydałoby się szersze wprowadzenie w temat oraz szersze wyjaśnienie znaczenia uzyskanych wyników. W szczególności, rozdział 4 pracy wygląda na nie zakończony, brak jest dyskusji wprowadzonych definicji oraz jakichś konkluzji.

**6. Jakie są słabe strony rozprawy i jej główne wady?**

a) Autor nie deklaruje w pracy, że ma doświadczenia praktyczne w zakresie jej przedmiotu, co rodzi przekonanie, że temat pracy powstał wyłącznie na podstawie czytanej literatury. Ponieważ literatura, szczególnie produkowana przez środowiska akademickie, jest często niewiarygodna, temat pracy nie ma dobrego umocowania pragmatycznego.
b) Wyniki pracy nie zostały wdrożone (co jest zrozumiałe), w związku z czym nie ma przesłanek do twierdzenia, że będą miały jakiekolwiek lub kiedykolwiek znaczenie praktyczne w zastosowaniach.
c) Praca nie zawiera odniesień do rzeczywistych potrzeb i preferencji podmiotu, dla którego są adresowane jej wyniki, czyli użytkownika kostki danych. Proponowane optymalizacje nie uwzględniają tych potrzeb i preferencji, w związku z czym można wątpić, czy w ogóle mają sens. Praca nie rozpatruje języka zapytań do kostki danych, związanych z tym preferencji użytkownika i ewentualnych pomocniczych struktur danych (w tym widoków) uwzględniających te preferencje.
d) Praca ogranicza struktury danych wspomagające proces ewaluacji zapytań do specyficznych widoków, pomijając mnóstwo innych struktur lub mechanizmów, które do tego celu mogłyby być wykorzystane. Można mieć wątpliwość, czy rzeczywiście widoki rozpatrywane w pracy mają aż tak zasadnicze znaczenie dla efektywności procesów decyzyjnych.
e) Praca sprawia wrażenie, że sprawą zasadniczej wagi dla jej autora było zbudowanie nietrywialnego modelu matematycznego i uzyskanie twierdzeń i wniosków z tego modelu, z pominięciem tego, czy ten model w sposób wiarygodny odwzorowuje rzeczywistość procesów decyzyjnych wykorzystujących kostki danych. Jest to pewien stereotyp „badań naukowych" preferowany przez niektóre środowiska (szczególnie uniwersyteckie wydziały matematyczno-informatyczne). Ten schemat prac badawczych, realizowany wcześniej w innych działach informatyki, udowodnił swoją praktyczną jałowość i zaowocował upadkiem autorytetu nauki akademickiej w środowiskach przemysłowych.

**7. Jaka jest przydatność rozprawy dla nauk technicznych?**

W dalszej perspektywie wyniki prezentowane w pracy mogłyby być przydane jako baza dla rozwoju naukowego oraz wzorzec dla porównań, przyszłych rozszerzeń i ewentualnych zastosowań. Bezpośrednia przydatność praktyczna wyników uzyskanych w rozprawie jest w mojej ocenie nieosiągalna.

**8. Do której z następujących kategorii Recenzent zalicza rozprawę?**

W mojej ocenie rozważałem wiele za i przeciw. Jako praktyk znający rzeczywistość przemysłową nigdy nie będę entuzjastą tego rodzaju prac. Jako naukowiec doceniam trud, który

autor wykonał celem zbudowania nietrywialnego modelu matematycznego oraz uzyskania twierdzeń i wniosków na bazie tego modelu. Wierzę, że będzie on istotny dla rozwoju naukowego zarówno autora jak i innych zainteresowanych osób. Moja konkluzja dotycząca rozprawy jest więc następująca:

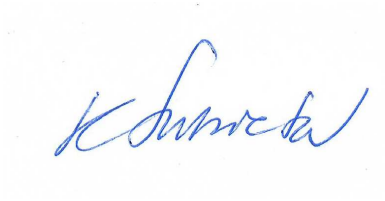
a/ ~~**nie spełniająca wymagań stawianych rozprawom doktorskim przez obowiązujące przepisy**~~

b/ ~~**wymagająca wprowadzenia poprawek i ponownego recenzowania**~~

**c/ Spełnia wymagania stawiane rozprawom doktorskim przez obowiązujące przepisy**

d/ ~~**spełniająca wymagania z wyraźnym nadmiarem**~~

e/ ~~**wybitnie dobra, zasługująca na wyróżnienie**~~



podpis