Łódź, 6.06.2019 r.

Dr hab. inż. Danuta Zakrzewska
Instytut Informatyki
Politechnika Łódzka

Recenzja rozprawy doktorskiej mgr Agnieszki Chądzyńskiej-Krasowskiej

pt. *„Wybrane metody ewaluacji i strojenia bazodanowych silników aproksymacyjnych bazujących na granularnych podsumowaniach danych"*

Recenzja niniejsza została przygotowana na prośbę Dziekana Wydziału Informatyki, Polsko-Japońskiej Akademii Technik Komputerowych (pismo z dnia 26.04.2019 roku).

1. Problematyka rozprawy

Tematyka rozprawy jest związana z problemem przetwarzania dużych zbiorów danych i dotyczy metod konstruowania przybliżonych zapytań. W dobie konieczności zarządzania ogromną liczbą danych, przy rosnących wymaganiach dotyczących czasu ich przetwarzania, problematyka ta jest ważna i aktualna. Jedną z obecnie stosowanych metod rozwiązania tego zagadnienia jest wykorzystanie przybliżonego przetwarzania zapytań, w którym działania są przeprowadzane na istotnie mniejszym zbiorze danych. Pojawia się tu jednak podstawowy problem dotyczący efektywności tego typu podejścia, a w szczególności dokładności wyników będących rezultatem działania przybliżonych kwerend. Rozwiązanie tak określonego zagadnienia stało się motywacją dla badań przeprowadzonych przez Doktorantkę.

W rozprawie Doktorantka skupiła się na metodach ewaluacji i na strojeniu bazodanowych silników aproksymacyjnych na przykładzie silnika działającego na granularnych posumowaniach danych. Autorka sformułowała dwie tezy pracy:

- *„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."*

- *„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".*

Zdefiniowany przez Doktorantkę problem badawczy, cel rozprawy oraz tezy zostały jasno i wystarczająco ogólnie sformułowane.

1. Treść rozprawy

Praca obejmuje 176 stron, składa się z 8 rozdziałów, wykazu bibliografii, dwóch dodatków oraz spisu tabel i rysunków. Treść rozprawy można opisać w sposób następujący:

Rozdział pierwszy zawiera wstępne wprowadzenie w tematykę rozprawy, wskazuje na aktualność tematu pracy oraz przedstawia kroki badawcze podjęte do realizacji celu pracy jak również tezy, których wykazanie prawdziwości stanowi główny cel badań prowadzonych przez Doktorantkę.

W rozdziale drugim wprowadzona została problematyka dotycząca przetwarzania dużych zbiorów danych. Przedstawione zostały metody tworzenia przybliżonych reprezentacji danych i możliwości ich wykorzystania do budowania zapytań. Rozpatrzony został silnik granularny w kontekście zastosowań dla histogramowych podsumowań danych.

Rozdział trzeci został poświęcony konstruowaniu miar dokładności przybliżonych wyników zapytań. Autorka rozważyła osiem miar, których własności zostały poddane analizie, ich znaczenie oceniono na podstawie opinii eksperckiej. Przedstawione zostały wyniki badań przeprowadzonych wśród użytkowników silników bazodanowych, których celem był wybór miary ocenionej jako najbardziej użyteczna. W efekcie zaproponowana została miara dokładności, która pozwala na ocenę dowolnej kwerendy analitycznej i jest pozytywnie postrzegana przez użytkowników.

W rozdziale czwartym Autorka opisała proces dostrajania parametrów silnika granularnego. Przedstawione zostało środowisko testowe oraz zbiory danych użyte w eksperymentach. Autorka rozważyła dobór parametrów silnika granularnego do konkretnych zbiorów. W szczególności, w tej partii rozprawy doktorantka stwierdziła, że zwiększenie paczki wierszy jest związane z „pierwiastkowo-liniowym" wzrostem budżetów przeznaczonych na

podsumowania danych. Pokazano również, że wstępne sortowanie danych pozytywnie wpływa na dokładność wyników zapytań

W kolejnym rozdziale rozważony został dobór parametrów w algorytmach podsumowujących. Celem badań było wybranie najbardziej efektywnych rozwiązań stosowanych w początkowej fazie granularyzacji danych. W efekcie Autorka wskazała histogram heurystyczny jako najbardziej adekwatny opis danych oryginalnych. Doktorantka zaproponowała również włączenie pojęć luki i wartości specjalnej do dziedziny atrybutu. Przeprowadzone zostały także badania dotyczące zwiększania efektywności metod budowania podsumowań danych w przypadku zależności międzykolumnowych.

Rozdział szósty dotyczy wykorzystania metod bazujących na podsumowaniach danych w zastosowaniach związanych z uczeniem maszynowym. W rozdziale tym skupiono się na zagadnieniach selekcji cech. Autorka wzięła tu pod uwagę miarę informacji wzajemnej. Doktorantka sformułowała i udowodniła twierdzenie, że informacja wzajemna wyliczana na podsumowaniach ma cechy analogiczne do informacji wzajemnej wyliczanej dla par kolumn. Zbadane zostały zmodyfikowane wersje znanych algorytmów selekcji cech działających na podsumowaniach danych. Przeprowadzono szereg eksperymentów, które pokazały efektywność metod selekcji cech działających na podsumowaniach danych.

W rozdziale siódmym Autorka zaproponowała model repozytorium metadanych granularnych, co pozwoliło na rozważenie kolejnego zastosowania metod bazujących na podsumowaniach – wizualizacji danych. Przeprowadzając szereg eksperymentów Doktorantka pokazała, że rozwiązania bazujące na granularnych podsumowaniach danych mogą stanowić podstawę do nowych technik wizualnej analizy danych.

Rozdział ósmy zawiera podsumowanie wyników badań prezentowanych w rozprawie.

W pracy zacytowano 105 pozycji literaturowych, w tym sześć publikacji z udziałem Autorki rozprawy.

Do pracy dołączono dwa dodatki. Pierwszy z nich zawiera przykładowe kwestionariusze ankiet dotyczących wyborów miar podobieństwa dla zapytań przybliżonych. W drugim dodatku umieszczony został kod SQL do wyznaczania przybliżonej wartości informacji wzajemnej.

2. Wyniki naukowe rozprawy

Do najważniejszych osiągniętych przez Autorkę wyników rozprawy należy zaliczyć:

- Zaproponowanie, opartego na miarach dokładności przybliżonych wyników zapytań, podejścia do ewaluacji bazodanowych silników aproksymacyjnych. Przeprowadzenie badań dotyczących najbardziej użytecznych miar ewaluacyjnych w kontekście potrzeb użytkowników. Opracowanie ankiety, umożliwiającej użytkownikom wskazanie miar, które będą w ich postrzeganiu najbardziej użyteczne przy ocenie efektów zapytań przybliżonych.
- Wprowadzenie modyfikacji algorytmów stosowanych w silniku bazodanowym jako efekt badań, przeprowadzonych w celu dostrajania parametrów silnika granularnego. Wskazanie parametrów zwiększających efektywność algorytmów podsumowujących.
- Opracowanie zmodyfikowanych wersji algorytmów uczenia maszynowego działających na podsumowaniach. Wskazanie metodyki badań efektywności algorytmów opartych na przybliżeniach w stosunku do ich pierwotnej wersji. Pokazanie efektywności metod selekcji cech opartych na podsumowaniach.
- Zaproponowanie modelu repozytorium metadanych granularnych, który może zostać wykorzystany do przybliżonego wykonywania zadań, których rozwiązania bazują na granularnych podsumowaniach danych. W szczególności mogą one stanowić podstawę do nowych technik wizualnej analizy danych.

Ważnym elementem rozprawy jest użyteczność otrzymanych wyników badań i możliwość praktycznego ich zastosowania. Doktorantka w swojej pracy wykorzystała opinie użytkowników dotyczące własności miar podobieństwa, korzystając z wypełnionego przez nich kwestionariusza ankiety. Jednak wybór przez Autorkę grup studentów studiów I i II stopnia kierunków związanych z Informatyką, jako reprezentatywnych użytkowników silników aproksymacyjnych nie wydaje się być w pełni uzasadniony. Autorka nie wyjaśniła motywacji takiego wyboru. Również konsultacje ustne z grupą ekspertów w dziedzinie zagrożeń sieciowych mogą nasuwać szereg wątpliwości. Nie zostały podane precyzyjne informacje na temat ich doboru, kompetencji oraz liczby. Warto jednak zwrócić uwagę, że pomimo zaistniałych wątpliwości inicjatywa Autorki potwierdzenia przez użytkowników wyboru miary podobieństwa jest cenna. W przyszłości warto byłoby jednak przeprowadzić badania biorąc pod uwagę alternatywne kryteria oceny.

Autorka przeprowadziła eksperymenty na istniejącym komercyjnie wdrożonym silniku granularnym, w którym podsumowania oparte są na histogramach. Przeprowadzone zostały

również badania nad zastosowaniem innych metod granularyzacji. Autorka skupiła się tu także na histogramach i wskazała histogram heurystyczny jako najbardziej adekwatną formę opisu danych. Wśród metod generowania reprezentacji przybliżonych Doktorantka wskazała również próbkowanie. Idea tej metody oraz strategie próbkowania zostały przez Autorkę opisane w Rozdziale 2.4. Jednak nie zostały one użyte w badaniach dla celów porównawczych. Wydaje się, że celowym byłoby, również i dla tych metod, sprawdzenie efektów zapytań przybliżonych przy uwzględnieniu rozważanej miary podobieństwa oraz ich porównanie z wynikami otrzymanymi przy zastosowaniu histogramów heurystycznych.

Wybór miar dokładności wyników zapytań nie został szczegółowo uzasadniony. Podane zostały jedynie kryteria ogólne. Nie zostały przez Autorkę wymienione miary, które zostały w wyniku testów odrzucone.

Doktorantka wykorzystała informację wzajemną wyliczaną na podsumowaniach do badań związanych z granularną selekcją cech. W tym celu pokazała, że własności informacji wzajemnej wyznaczanej na podsumowaniach są takie same jak własności klasycznej informacji wzajemnej. Dowody Twierdzenia 6.1 oraz Twierdzenia 6.2 są analogiczne. Stąd podawanie ich obydwu, jest nadmiarowe, zwłaszcza że praca głównie oparta jest na eksperymentach.

3. Ocena redakcji rozprawy

Praca została zorganizowana w sposób dość przejrzysty. Rozdziały rozpoczynają się krótkim wstępem wprowadzającym w omawianą tematykę. Na końcu rozdziałów umieszczone zostały podsumowania, co pozwala na uporządkowanie przeczytanej treści. Autorka nie zamieściła jednak rozdziału wprowadzającego, w którym zdefiniowane byłyby użyte w pracy pojęcia i oznaczenia. Ułatwiłoby to czytanie pracy. Nie został również wyróżniony rozdział zawierający przegląd literatury, stąd trudność może stanowić wyróżnienie zakresu wprowadzonych przez Doktorantkę rozwiązań. Praca zawiera szereg tabel i rysunków, co znacząco ułatwia jej czytanie. Dodatkowo Autorka zamieściła wiele przykładów, które ułatwiają zrozumienie treści, choć nie zostały wprowadzone w ustrukturyzowany sposób. Język rozprawy jest dość poprawny, jednak w wielu miejscach Autorka używa kolokwializmów. Występują również błędy językowe. Uwagi dotyczące języka zamieszczę w następnej sekcji, wraz z innymi uwagami szczegółowymi do rozprawy.

4. Uwagi szczegółowe do rozprawy

W szczególności:

- „bazujących ma" powinno być „bazujących na" (str 13 i 17)
- „ słusznej miary" nie zostało zdefiniowane pojęcie „słusznej" miary (str. 17, 44, 48)
- „średniej... na 10 TB danych" powinno być „średniej ...dla 10 TB danych (str 19)
- Występuje redundancja dla tekstu umieszczonego we Wstępie (Rozdział 1) oraz w Rozdziale 2. Dotyczy to np. stwierdzenia dotyczącego prawa Moore'a
- Nie jest jasne stwierdzenie „Falki są dobre w wychwytywaniu cech zbioru.." (str. 23). Własność „dobre" nie została wcześniej zdefiniowana.
- W Rozdziale 2. Autorka wymienia różne rodzaje podsumowań. Jednak nie wszystkie zostały w pracy zdefiniowane (dotyczy to np. „ε-sieci")
- Nie jest jasne zdanie dotyczące próbkowania: „W przypadku ogólnym, obejmującym między innymi podzapytania, sprawa nie jest już taka prosta, ale częściowo rozwiązywalna w proceduralny sposób" (str 22).
- Wyjaśnienia wymaga stwierdzenie „...różnorodność metodologii estymacyjnych opartych na próbkowaniu jest oszałamiająca" (str 24)
- Autorka używa sformułowania „naiwne podejście" przy opisie histogramów równej głębokości (str 29). Jednak pojęcie to nie zostało wyjaśnione.
- Nie jest jasne zdanie dotyczące histogramów: „cierpią one na pewien rodzaj przekleństwa wymiaru" (str 35)
- Jest „bazujące ma pojęciu" powinno być „bazujące na pojęciu" (str 43)
- Jest „powinna zgodna" zamiast „powinna być zgodna" (str 48)
- Nie jest jasne czy „dziesięciokrotność paczki" jest równoznaczne z pojęciem „dziesięciopaczki" (str 96)
- Pojęcia „zakres" i „dopełnienie" użyte we wzorze (6.3) zostały wyjaśnione jedynie w podpisie do Rys. 6.1
- We wzorze opisującym informację wzajemną (6,1) znajduje się logarytm przy podstawie 10 (log), natomiast dowody twierdzeń 6.1 i 6.2 zostały przeprowadzone dla logarytmu naturalnego

- Przedstawione w rozprawie algorytmy są oznaczane za pomocą angielskiego słowa „Algorithm", natomiast w treści Autorka odwołuje się do nich za pomocą polskiego odpowiednika Algorytm
- Brak jest jednolitego formatu pozycji literaturowych

5. Konkluzja

Stwierdzam, że zaprezentowana rozprawa doktorska pani mgr Agnieszki Chądzyńskiej -Krasowskiej stanowi oryginalne rozwiązanie problemu naukowego. W pracy Doktorantka wykazała głęboką wiedzę zarówno z zakresu przetwarzania dużych zbiorów danych jak i ich analizy. Realizacja pracy pokazała umiejętność Doktorantki do samodzielnego prowadzenia badań naukowych. Reasumując stwierdzam, że rozprawa pt. „ „*Wybrane metody ewaluacji i strojenia bazodanowych silników aproksymacyjnych bazujących na granularnych podsumowaniach danych*" spełnia wymagania stawiane pracom doktorskim, określone w Ustawie o stopniach naukowych i tytule naukowym. Dorobek publikacyjny Doktorantki związany z powstawaniem rozprawy w postaci materiałów konferencji międzynarodowych uważam za wystarczający.

W związku z powyższym wnioskuję o dopuszczenie rozprawy do publicznej obrony.

Danuta Zakrzewska