RDITI-wpł. 10.10.2022
M. Skwarek

Poznań, 30.09.2022

dr hab. inż. Robert Wrembel, prof. nadzw.
Politechnika Poznańska
Wydział Informatyki i Telekomunikacji
Instytut Informatyki
ul. Piotrowo 2
60-965 Poznań
e-mail: robert.wrembel@cs.put.poznan.pl

Recenzja rozprawy doktorskiej

mgr inż. Krzysztofa Pasternaka

pt. Strumieniowe hurtownie danych zorientowane na przetwarzanie wielkich zbiorów danych kontekstowych

1. Tematyka i zarys problemu

Recenzowana rozprawa doktorska mgr inż. Krzysztofa Pasternaka jest poświęcona architekurze strumieniowej hurtowni danych.

Przetwarzanie danych strumieniowych w celu np.: (1) wykrywania w nich wzorców lub anomalii, (2) korelowania wartości pojawiających się w równolegle przetwarzanych strumieniach, (3) wykrywania wzorców, (4) obliczania zagregowanych wartości w zadanym oknie czasowym, jest jednym w ważniejszych nurtów badawczych w zakresie nowoczesnych architektur przetwarzania danych. Wyniki badawcze mają bezpośrednie przełożenie na praktykę, np. w systemach wykrywania nieautoryzowanego użycia kart płatniczych.

Wyniki prac badawczych w zakresie przetwarzania danych strumieniowych są od lat prezentowane na najlepszych światowych konferencjach i w czasopismach. Przykładowo, serwis DBLP udostępnia prawie 9 tys. artykułów dotyczących 'data stream'. W tym kontekście, tematyka rozprawy dotyczy problemów w jednym z ważniejszych nurtów badawczych.

Motywacją do podjęcia zagadnień adresowanych w rozprawie jest rzeczywisty problem zarządzania dystrybucją paliw płynnych. W szczególności, problem dotyczy składowania danych strumieniowych w repozytorium, z możliwością bieżącego uaktualniania zawartości tego repozytorium, możliwością jego przeszukiwania i agregowania danych strumieniowych. Repozytorium takie w rozprawie nazwano strumieniową hurtownią danych.

2. Struktura rozprawy

Recenzowana rozprawa doktorska składa się z 10-ciu rozdziałów. Rozdziały 1-4 stanowią wprowadzenie do problematyki zarządzania i dystrybucji paliw płynnych oraz opis stanu wiedzy. Rozdziały 5-9 opisują wyniki rozprawy. Rozdział 5 omawia dane kontekstowe, które zostały wprowadzone do rozprawy jako nowy element. Rozdział 6 omawia ogólny zarys architektury strumieniowej hurtowni danych i modele logiczne reprezentowania danych kontekstowych. Rozdział 7 prezentuje opracowane przez Doktoranta: tzw. silnik strumieniowej kostki CUBIT i tzw. wielowymiarowy bitowy indeks zakresowy. Rozdział 8 prezentuje trzy

autorskie algorytmy stronicowania dla strumieniowyej hurtowni danych. Rozdział 9 zawiera ocenę eksperymentalną zaproponowanych algorytmów. Rozdział 10 podsumowuje rozprawę.

3. Wyniki rozprawy

Do najważniejszych osiągnięć rozprawy zaliczam:

- silnik strumieniowej kostki CUBIT,
- wielowymiarowy bitowy indeks zakresowy,
- algorytmy stronicowania, wraz z oceną ich podstawowych charakterystyk.

Każda z powyższych koncepcji została opracowana teoretycznie i zaimplementowana.

4. Uwagi

Poniżej przedstawiam uwagi o charakterze dyskusyjnym, które nie umniejszają wartości wyników osiągniętych w recenzowanej rozprawie.

1. Na stronie 92 Doktorant stwierdza: "Powyższe rozważania, razem z przedstawionymi w dalszej części rozdziału treściami, stanowią zarazem dowód tezy". Teza natomiast brzmi: "możliwe jest zaprojektowanie strumieniowej hurtowni danych zorientowanej na przetwarzanie wielkich zbiorów danych kontekstowych". W celu udowodnienia tezy należy zaprojektować, zbudować i ocenić eksperymentalnie pod kątem wskazanych kryteriów strumieniową hurtownię danych. Rozważania teoretyczne tego rozdziału nie dowodzą tezy. Dowód ten możemy znaleźć znacznie później, pod koniec rozprawy.
2. W punkcie 6.2.1 przedstawiono model logiczny hurtowni danych strumieniowych. Nie jest jasne co jest faktem w tym modelu; czy jest nim pojedynczy pomiar, czy szereg czasowy; jeśli szereg czasowy to o stałej, czy zmiennej długości? W schemacie przedstawionym na rys. 6.2 nie jest jasne jakie wartości przyjmują atrybuty *schema_attributes* i *schema_types*.
3. Na stronie 96 stwierdzono: "... tabele wymiarów w obu modelach mogą być fizycznie tymi samymi tabelami lub istnieć niezależnie. Pierwsze rozwiązanie zapewnia spójność danych oraz jest mniej kosztowne pamięciowo, ale wymusza przechowywanie strumieni danych oraz metadanych w tej samej bazie. W drugim rozwiązaniu konieczne jest zadbanie o poprawne replikacje danych, ale możliwe jest zastosowanie odrębnego magazynu danych dla metadanych". Dlaczego pierwsze rozwiązanie jest mniej kosztowne pamięciowo, skoro w obu rozwiązaniach przechowujemy tyle samo danych? Jaka jest zaleta drugiego rozwiązania?
4. W punkcie 7.1.1 poruszono problem rekonstruowania zarchiwizowanych strumieni. Jaka jest wydajność takiej rekonstrukcji?
5. W rozdziale 7 omówiono zagadnienie agregowania strumieni danych. Jak się domyślam także w celu zmniejszenia wolumenu danych. Alternatywnym rozwiązaniem jest kompresowanie szeregów czasowych. Czy kompresja nie byłaby lepsza w tym przypadku?
6. W rozdziale 7 omówiono także autorską strukturę indeksową. Czy opracowano dla niej algorytm uaktualniania?
7. W rozdziale 9 zaproponowano metryki jakości usług dla strumieniowej hurtowni danych, tj. jakość usług konsumenta i jakość usług producenta. Dlaczego standardowe miary przepustowość i czas odpowiedzi nie były wystarczające?
8. W rozdziale 9 przedstawiono także wyniki oceny eksperymentalnej autorskich algorytmów. Czy dokonano tego za pomocą symulacji komputerowej? Zabrakło omówienia danych podstawowych dotyczących zmiennych eksperymentów takich jak:

obciążenie zapytaniami, obciążenie strumieniami, wolumen przetwarzanych danych, selektywność zapytań.

5. Ocena końcowa

Podsumowując, uważam, że cel rozprawy został osiągnięty. Mgr inż. Krzysztof Pasternak wykazał, że zastosowanie zaproponowanych w rozprawie technik gromadzenia, zarządzania i analizowania danych strumieniowych spełnia przyjęte w rozprawie założenia.

Sposób formułowania problemów badawczych i metodyka ich rozwiązania jest właściwa dla badań naukowych w informatyce. Zastosowana metodyka bazuje na analizie teoretycznej problemu, budowie modelu matematycznego, implementacji rozwiązania i jego ocenie eksperymentalnej w postaci symulacji.

Stosowalność opracowanych w ramach rozprawy rozwiązań została uwiarygodniona trzema krajowymi zgłoszeniami patentowymi i jednym liczącym się artykułem za 140pkt (wg. listy MNiSzW). Wyniki badań opublikowano także kilku niepunktowanych lub nisko-punktowanych czasopismach i konferencjach. Fakt uzyskania trzech krajowych patentów bazujących na wynikach rozprawy, wskazuje, że rozprawa może mieć pozytywny efekt aplikacyjny w przemyśle.

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