wpł. RDITT
16.09.2022

Dr hab. Zygmunt Mazur, prof. uczelni
Politechnika Wrocławska
Wydz. Informatyki i Telekomunikacji
zygmunt.mazur@pwr.edu.pl

Wrocław, 15 września 2022 r.

Recenzja rozprawy doktorskiej

mgr inż. Krzysztofa Pasteraka

zatytułowanej

Strumieniowe hurtownie danych zorientowane na przetwarzanie wielkich zbiorów danych kontekstowych

Promotor rozprawy: prof. dr hab. inż. Marcin Gorawski
Dziedzina: nauki inżynieryjno - techniczne
Dyscyplina: informatyka techniczna i telekomunikacja

1. Problem badawczy w rozprawie i jego znaczenie

1.1. Cel, zakres i charakter rozprawy

Skutkiem rozwoju technologicznego jest powstanie urządzeń pomiarowych i usług, które umożliwiają niemal nieprzerwane monitorowanie świata rzeczywistego, a w konsekwencji ciągłą detekcję anomalii w strumieniach danych. W rozprawie doktorskiej przedstawiony został rzeczywisty system detekcji wycieków paliw płynnych na stacjach paliw płynnych. System detekcji wycieków paliwa jest wynikiem autorskich badań naukowych prowadzonych w zespole badawczym, kierowanym przez promotora we współpracy z partnerem biznesowym z sektora paliw płynnych.

Przykładem anomalii krytycznej jest wyciek paliwa ze zbiornika na stacji paliw płynnych, stanowiącym zdarzenie zachodzące w świecie rzeczywistym, które wymaga monitorowania i natychmiastowej reakcji w momencie wykrycia - przeprowadzania działań diagnostyczno-serwisowych lub naprawczych. Występowanie anomalii krytycznych jest nieuniknione, a szybkość ich detekcji jest kluczowa ze względu na ich negatywny wpływ na środowisko, sytuację finansową oraz bezpieczeństwo i pewność diagnostyki obiektu, w którym taka anomalia wystąpiła.

Potrzeba analizy danych historycznych narzuca konieczność zastosowania systemów zaawansowanych hurtowni danych. W rozprawie zaproponowana jest sugestia, by w procesie projektowania systemów zaawansowanych hurtowni danych detekcja anomalii krytycznych stanowiła jedną z głównych motywacji i wyznaczników dotyczących metod ekstrakcji i magazynowania danych, a nie jedynie etap analizy eksploracyjnej.

W rozprawie anomalia zdefiniowana jest jako zdarzenie występujące w świecie rzeczywistym, które odbiega od przyjętej normy lub specyfikacji. Występujące anomalie w świecie rzeczywistym stanowią zdarzenia, których reprezentacji szuka się w danych pomiarowych. Stąd detekcja anomalii w trybie on-line wymaga zdefiniowania natury zdarzeń, ich charakterystyk, trybu występowania oraz zależności między nimi. Problem detekcji anomalii w trybie on-line nabiera znaczenia zwłaszcza w kontekście rozwoju technologicznego, który stał się źródłem danych opisujących szereg aspektów życia codziennego oraz procesy biznesowe i przemysłowe.

Zasadnicza tematyka rozprawy doktorskiej dotyczy problemów badawczych związanych ze strumieniowymi hurtowniami danych kontekstowych. **Tematyka rozprawy w pełni wpisuje się w zakres dziedziny nauk inżynieryjno - technicznych, w dyscyplinie informatyka techniczna i telekomunikacja.**

Celem rozprawy doktorskiej jest przedstawienie powiązanych zagadnień teoretycznych, przedstawienie propozycji nowych metod i modeli, przeprowadzenie badan eksperymentalnych oraz przeprowadzenie analizy ich wyników. W rozprawie doktorant zaprezentował modele i metody przetwarzana strumieni danych w hurtowniach danych oraz aspekty ich stosowania w systemach wykrywających anomalie. Badania przedstawione w rozprawie zostały opracowane na podstawie opublikowanych współautorskich prac naukowych oraz zgłoszeń patentowych.

1.2. Tezy rozprawy

Pierwszym zagadnieniem omawianym w rozprawie jest system dystrybucji i składowania paliw płynnych, w którym przeprowadzono badania nad problemem detekcji anomalii krytycznych i wykrycie zdarzeń od przyjętej normy. Wynikiem tych badan jest opracowanie metody wykrywania wycieków wykorzystującej detekcje i interpretacje trendów (algorytm TUBE), a wniosek z tej części badań to spostrzeżenie o konieczności wzięcia pod uwagę kontekstu badanych zjawisk przy ich analizie. Na podstawie tego wniosku doktorant sformułował postać pierwszej tezy w rozprawie:

Teza 1

„Uzyskanie w pełni jednoznacznych wyników analizy danych ukierunkowanej na wykrywanie zdarzeń anomalnych jest możliwe dopiero po uwzględnieniu kontekstu występowania poszczególnych anomalii, na który składają się dane współistniejące w czasie i przestrzeni oraz powiązane semantycznie z analizowanym zjawiskiem".

Wnioski z tej części badan posłużyły za podstawę do sformułowania teorii danych kontekstowych, wraz z podaniem ich definicji, klasyfikacji oraz zarysu metod ich przetwarzania. Ten obszar tematyczny stanowi drugie istotne zagadnienie recenzowanej rozprawy. Z tego zagadnienia wynika bezpośrednio kolejne, stanowiące próbę praktycznego ujęcia tematu przetwarzania danych kontekstowych: model strumieniowej hurtowni danych kontekstowych. Został on zaprezentowany jako kompletny system składowania i przetwarzania danych kontekstowych, ukierunkowany na wykrywanie oraz weryfikacje anomalii krytycznych. W tym zakresie przedstawiano drugą tezę rozprawy:

Teza 2

„Możliwe jest zaprojektowanie strumieniowej hurtowni danych zorientowanej na przetwarzanie wielkich zbiorów danych kontekstowych, wykorzystującej wielotorowy model przetwarzania danych, w którym analiza danych krytycznych jest wsparta przez przeprowadzana niezależnie wieloaspektową analizę danych kontekstowych, w celu uwiarygodnienia wyników tej pierwszej".

W ramach modelu strumieniowej hurtowni danych kontekstowych zaproponowano i opisano szereg metod i modeli, przeznaczonych do wspierania przetwarzania danych kontekstowych. Są to w szczególności: silnik strumieniowej kostki CUBIT oraz indeks przestrzenny BRI. Ta pierwsza jest odpowiednikiem kostki OLAP dla wielowymiarowych danych strumieniowych. Drugie rozwiązanie to wielowymiarowy bitowy indeks zakresowy, wspierający wykonywanie zapytań o agregaty zakresowe w wielowymiarowej przestrzeni cech.

Ostatnim zagadnieniem omawianym w rozprawie jest problem efektywnego dostarczania agregatów wielowymiarowych. W ramach tego problemu zaprojektowano trzy nowe adaptacyjne algorytmy stronicowania, przeznaczone dla omówionego silnika CUBIT. Algorytmy te wykorzystują metody optymalizacji wielokryterialnej do zapewnienia należytej jakości usług, zarówno klienta (użytkownika), jak i źródła (bazy) danych. Nowe algorytmy zostały poddane analizie weryfikacyjnej oraz porównawczej – zarówno pomiędzy sobą, jak i z poprzednia generacja algorytmów. Badania przeprowadzono przy użyciu dwóch zaproponowanych metryk jakości usług dla strumieniowych hurtowni danych. Zaproponowane algorytmy i metryki zostały ujęte w formie trzeciej tezy rozprawy:

Teza 3

„Zastosowanie metod optymalizacji wielokryterialnej w procesie stronicowania w strumieniowych hurtowniach danych oraz uwzględnienie bieżących parametrów pracy i istniejących ograniczeń, pozwala na zwiększenie jakości usług, rozumianej zarówno jako poprawę efektywności i ciągłości dostarczania danych użytkownikowi, jak również zmniejszenie obciążenia źródła danych".

Weryfikacje tez rozprawy przeprowadzono w sposób teoretyczny oraz empiryczny. Pierwszy sposób polegał na budowie modeli i analizie powiązanych zagadnień teoretycznych. Sposób empiryczny opierał się na wynikach eksperymentów oraz wnioskach sformułowanych przy pracy z rzeczywistymi obiektami przemysłowymi.

Na weryfikacje pierwszej tezy rozprawy składało się scharakteryzowanie sieci stacji paliw i zachodzących w niej procesów, w tym anomalii krytycznych oraz dyskusja na temat jakości wyników algorytmu wykrywania wycieków paliwa w świetle współistnienia innych zjawisk, stanowiących kontekst dla tych wycieków.

Druga teza została zweryfikowana przez sformułowanie modelu strumieniowej hurtowni danych kontekstowych, wliczając w to zarówno modele cząstkowe poszczególnych baz danych, jak i modele silnika CUBIT oraz indeksu BRI.

Weryfikacja trzeciej tezy została przeprowadzona w sposób empiryczny, przez wykonanie szeregu eksperymentów porównawczych dla zaproponowanych trzech nowych algorytmów wypełniania stron, przy jednoczesnym wykorzystaniu dwóch nowych metryk jakości usług.

W zakończeniu rozprawy wskazano na otwarte problemy badawcze, wliczając w to: rozwój modelu silnika CUBIT oraz indeksu BRI, a także udoskonalenie zaproponowanych algorytmów wypełniania stron. Ponadto, opisano również potencjalne drogi rozwoju poszczególnych metod i modeli, takie jak: adaptacje istniejących rozwiązań do innych problemów świata rzeczywistego oraz integracje zaprezentowanego modelu strumieniowej hurtowni danych kontekstowych z modelem Spichlerza Agregatów.

1.3. Zawartość rozprawy

Rozprawa doktorska liczy 220 stron i tworzy ją 10 rozdziałów. W rozdziale 1 przedstawiono tezy, cel i zakres rozprawy. W rozdziale 2 przedstawiono opisy ewolucji modeli hurtowni danych, skupiając się przede wszystkim na strumieniowych hurtowniach danych, jako tło dla kolejnych zagadnień poruszanych w rozprawie. Dokonano obszernego przeglądu literatury oraz przedstawiono historie prac badawczych nad Materializowana Lista Agregatów (MAL), która jako jeden z głównych komponentów modelu strumieniowej hurtowni danych, stała się inspiracja do opracowania modelu silnika strumieniowej kostki CUBIT.

Rozdział trzeci rozprawy przedstawia opis modelu sieci stacji paliw, jako źródła strumieni danych. Dokonano klasyfikacji danych paliwowych oraz omówiono podstawowe zjawiska fizyczne zachodzące na stacjach paliw. Zawarto opis modelu konceptualnego przykładowej hurtowni danych paliwowych – jego celem jest zwrócenie uwagi na istotne aspekty analizy takich danych oraz ich wielowymiarowość.

Rozdział czwarty rozprawy opisuje anomalie występujące w procesie dystrybucji i składowania paliw płynnych. Przedstawiona została klasyfikacja anomalii oraz opis metody wykrywania wycieków paliwa. Metoda ta stanowi jedna z kontrybucji pobocznych rozprawy. Przedstawiono streszczenie w formie graficznej wyników ewaluacji tej metody i sformułowano wnioski, stanowiące kierunki do dalszych badań w zakresie analizy danych kontekstowych.

Rozdział piąty rozprawy jest pierwszym z rozdziałów zawierających opis teoretycznego modelu danych kontekstowych. Podano definicje formalne danych

kontekstowych, dokonano ich klasyfikacji oraz opisano modele przetwarzania poszczególnych ich rodzajów. Zdefiniowano siedem poziomów kontekstowości, które określają stopień uwzględniania danych kontekstowych w obliczeniach.

Rozdział szósty rozprawy przedstawia propozycje modelu strumieniowej hurtowni danych kontekstowych, będącej wielotorowym systemem składowania i przetwarzania danych na różnych poziomach kontekstowości. Omówiono ogólny zarys architektury, a także zaprezentowano modele logiczne trzech rodzajów baz danych dla trzech rodzajów danych kontekstowych – bazy te wchodzą w skład przytoczonego modelu strumieniowej hurtowni danych kontekstowych, stanowiąc dla niego warstwę składowania danych.

Rozdział siódmy rozprawy omawia wybrane modele procesów, związanych z funkcjonowaniem strumieniowego serwera OLAP, istotnych dla zaproponowanej w rozprawie strumieniowej hurtowni danych kontekstowych. Są to w szczególności: silnik strumieniowej kostki CUBIT, będący rozwinięciem koncepcji silnika Materializowanej Listy Agregatów oraz wielowymiarowy bitowy indeks zakresowy BRI.

Rozdział ósmy zawiera propozycje trzech nowych algorytmów stronicowania dla strumieniowych hurtowni danych, które cechują się adaptacyjnością, elastycznością i przewidywalnością w stosunku do swoich poprzedników. Algorytmy te wykorzystują metody optymalizacji wielokryterialnej. Przedstawiono obszerny opis procesu projektowania tych algorytmów, uwzgledniający między innymi definicje przestrzeni rozwiązań, funkcji celu oraz strategii wyboru rozwiązań niezdominowanych.

Rozdział dziewiąty rozprawy wprowadza dwie nowe metryki jakości usług dla strumieniowych hurtowni danych. Są to: jakość usług konsumenta, rozumiana jako efektywność dostarczania nowych danych użytkownikowi oraz jakość usług producenta, rozumiana jako stabilność funkcjonowania źródła danych. Ponadto, rozdział ten zawiera obszerne sprawozdanie z wykonanych badań eksperymentalnych.

Dziesiąty rozdział rozprawy, stanowi podsumowanie. Zawiera wnioski z ogółu wykonanych prac badawczych, wskazuje plany rozwoju i nowe perspektywy badawcze.

Bibliografia rozprawy liczy 132 pozycje literaturowe, zamieszczono spisy 74 rysunków i 6 tabel. Do pracy dołączony jest pendrive, który zawiera dwa pliki: z pracą doktorską w formacie PDF oraz krótkie dwu stronicowe streszczenie rozprawy w języku polskim.

2. Wkład naukowy autora rozprawy

2.1. Poprawność i oryginalność postawionych tez

Doktorant w rozprawie doktorskiej dokonał analizy zagadnień teoretycznych oraz praktycznych związanych z modelem strumieniowej hurtowni danych zorientowanej na przetwarzanie wielkich zbiorów danych kontekstowych. W trakcie realizacji tego celu doktorant dokonał weryfikacji tez postawionych w rozprawie.

Tezy dotyczyły trzech zagadnień:

- uznania istotności danych kontekstowych w analizie ukierunkowanej na wykrywanie zjawisk niepożądanych;

- weryfikacji możliwości stworzenia kompletnego systemu przeznaczonego do składowania i przetwarzania danych kontekstowych;
- potwierdzenia zasadności stosowania wybranych technik optymalizacyjnych w procesach transmisji danych, jako metod zwiększania jakości usług.

Weryfikacje tez rozprawy doktorant przeprowadził w sposób teoretyczny oraz empiryczny. Weryfikacje teoretyczne zrealizowano poprzez budowę modeli i analizie powiązanych zagadnień teoretycznych. Weryfikacja empiryczna opierała się na wynikach przeprowadzonych eksperymentów oraz na wnioskach sformułowanych przy pracy z rzeczywistymi obiektami przemysłowymi. Punktem wyjścia do obu podejść był przykład systemu dystrybucji i składowania paliw płynnych. Głównym problemem badawczym, a zarazem bezpośrednią motywacja do podjęcia tematyki rozprawy, było wykrywanie wycieków z podziemnych zbiorników paliw.

Postawiona pierwsza teza rozprawy (teza 1) zakładała, że uzyskanie w pełni jednoznacznych wyników detekcji anomalii jest możliwe dopiero po uwzględnieniu kontekstu występowania poszczególnych zjawisk. Kontekst ten opisany powinien być przez dane współistniejące w czasie i przestrzeni oraz powiązane semantycznie z analizowanym zjawiskiem. W tym zakresie, teza ta odnosiła się do teoretycznych aspektów danych kontekstowych, podkreślając ich istotność przy analizie danych zasadniczych.

Weryfikację tezy 1 doktorant rozpoczął od scharakteryzowania sieci stacji paliw i zachodzących w niej procesów. Dokładne poznanie charakterystyki zjawisk zachodzących na stacjach paliw pozwoliło na identyfikacje podstawowych zjawisk niepożądanych oraz na dokonanie ich klasyfikacji pod katem lokalności i istotności. Przedstawiono również opracowana metodę wykrywania wycieków paliwa (algorytm TUBE) oraz uzyskane przez nią wyniki. Dyskusja na temat jakości otrzymanych wyników, a także wnioski wynikłe z pracy z rzeczywistymi zbiorami danych, pozwoliły na sformułowanie podstaw analizy kontekstowej, udowadniając zarazem tezę 1.

Druga teza rozprawy zakładała, że możliwe jest zaprojektowanie strumieniowej hurtowni danych zorientowanej na przetwarzanie wielkich zbiorów danych kontekstowych. Hurtownia ta powinna wykorzystywać wielotorowy model przetwarzania danych, w którym analiza danych krytycznych byłaby wsparta przez przeprowadzaną niezależnie wieloaspektowa analizę danych kontekstowych. W ten sposób sformułowana, teza ta odnosiła się do praktycznego wykorzystania danych kontekstowych. Zasadność postawienia tej tezy wynikała bezpośrednio ze sformułowania i potwierdzenia tezy 1.

Weryfikacja tezy 2 została rozpoczęta od przybliżenia klasy zaawansowanych hurtowni danych, ze szczególnym uwzględnieniem strumieniowych hurtowni danych (rozdział 2). Kolejnym etapem prac badawczych było sformułowanie podstaw teoretycznych danych kontekstowych, włączając w to ich model, klasyfikację oraz metody przetwarzania (rozdział 5). Na bazie tych rozważań, zaprojektowano oraz w

szczegółach opisano (rozdział 6) model strumieniowej hurtowni danych kontekstowych (CtxDW). Hurtownia ta jest rozwinięciem koncepcji strumieniowej hurtowni danych poprzez jej adaptacje do wielotorowego przetwarzania danych kontekstowych. Ostatnią składową weryfikacji omawianej tezy było opracowanie modelu silnika CUBIT oraz koncepcji wielowymiarowego bitowego indeksu zakresowego BRI, jako kluczowych elementów strumieniowego serwera OLAP (rozdział 7). Pierwsze rozwiązanie stanowi strumieniową adaptację kostki OLAP i zostało opracowane jako rozwinięcie silnika Materializowanej Listy Agregatów (MAL). Drugie rozwiązanie jest propozycją wysuniętą w stronę efektywnego obliczania wielowymiarowych agregatów na danych ciągłych, wykorzystując do tego równoległość na poziomie bitów i danych. Całość przedstawionych w wymienionych rozdziałach metod i narzędzi składa się zarazem na dowód poprawności tezy 2.

Trzecia teza rozprawy (teza 3) zakładała, że zastosowanie metod optymalizacji wielokryterialnej oraz uwzględnienie bieżących parametrów pracy i ograniczeń wpłynie pozytywnie na jakość usług. Jakość ta powinna być rozumiana nie tylko jako efektywność dostarczania danych użytkownikowi, ale również jako płynność pracy źródła danych. W tym zakresie, teza ta odnosiła się do funkcjonowania strumieniowej hurtowni danych kontekstowych – jej postawienie implikowane było sformułowaniem i weryfikacja tezy 2.

Weryfikacja tezy 3 została rozpoczęta od silnika CUBIT, a dokładniej od będącego integralna jego częścią algorytmu wypełniania stron. Algorytm ten rozwiązuje problem optymalizacyjny o dwóch przeciwstawnych celach: minimalizacji opóźnienia w dostarczaniu wyników zapytań użytkownikom oraz minimalizacji obciążenia bazy danych. Przeprowadzono analizę (rozdział 8) postawionego problemu optymalizacyjnego oraz sformułowano ogólny zarys algorytmu oraz jego trzy warianty: TRAFF, LATEN i HYBRI, zwane dalej algorytmami wypełniania stron. Algorytmy te zostały zaprojektowane w celu zastąpienia poprzedniej generacji algorytmów wypełniania stron (stosowanych w silniku MAL). Wszystkie algorytmy zostały zaimplementowane w postaci symulatora zdarzeń dyskretnych i poddane badaniom eksperymentalnym, których wyniki zaprezentowano w rozdziale 9. Zaproponowano dwie nowe metryki jakości usług: konsumenta i producenta, a otrzymane wyniki potwierdziły słuszność postawionych założeń i dowiodły prawdziwości tezy 3.

2.2. Wkład autora w rozwój naukowy dyscypliny

Wyniki naukowe zaprezentowane w rozprawie doktorskiej stanowią duży wkład doktoranta w rozwój dyscypliny naukowej. Otrzymane wyniki można zaprezentować w formie wyników cząstkowych zdefiniowanych ośmiu zadań, które można podzielić na dwie grupy: zadania główne (podstawowe) i pomocnicze.

Lista oryginalnych wyników badań:

1. **Pomocnicze zadanie A**

Opracowanie modelu sieci stacji paliw, jako złożonego źródła danych krytycznych i kontekstowych oraz klasyfikacja anomalii paliwowych (rozdziały 3 i 4), model sieci stacji oraz klasyfikacja anomalii stanowią podstawę do analizy procesów zachodzących na stacjach paliw, a dodatkowo stanowią podstawę do dalszego rozwoju metod wykrywania anomalii paliwowych [38, 28];

2. **Pomocnicze zadanie B**

Opracowanie algorytm TUBE jako metoda analizy danych bieżących, ukierunkowana na wykrywanie anomalii krytycznych (rozdział 4). Opis algorytmu TUBE został opublikowany w czasopiśmie naukowym z listy filadelfijskiej [39], a jego wyniki posłużyły za motywacje do podjęcia badan nad analizą danych kontekstowych (patenty [70, 40, 71]);

3. **Podstawowe zadanie C**

Opracowanie teoretycznego modelu danych kontekstowych, uwzgledniający ich klasyfikacje oraz hierarchie poziomów przetwarzania (rozdział 5). Model wprowadza pojęcie danych kontekstowych oraz krytycznych, podział tych pierwszych na dane kontekstowe historyczne, przestrzenne i środowiskowe, a następnie definiuje siedem poziomów kontekstowości tworzących hierarchie. Model powstał w oparciu o wyniki wcześniejszych prac badawczych, przedstawionych w artykułach naukowych [66, 36, 37];

4. **Podstawowe zadanie D**

Model strumieniowej hurtowni danych kontekstowych, będącej rozwinięciem koncepcji strumieniowej hurtowni danych poprzez jej ukierunkowanie na przechowywanie i przetwarzanie danych kontekstowych. Opisany (rozdział 6) model wykorzystuje wielotorową architekturę, w której każdy tor odpowiada za przetwarzanie innego typu danych kontekstowych, wliczając w to również tor krytyczny, dedykowany danym bieżącym [66, 36, 37];

5. **Podstawowe zadanie E**

Model silnika CUBIT, stanowiącego adaptację koncepcji kostki danych OLAP na potrzeby wielowymiarowego przetwarzania danych strumieniowych (rozdział 7). Model stanowi rozwiniecie silnika Materializowanej Listy Agregatów (MAL) poprzez uwzględnienie wielowymiarowości danych, przez co staje się strumieniową adaptacją kostki OLAP [67, 68];

6. **Podstawowe zadanie G**

Model indeksu BRI, czyli wielowymiarowego bitowego indeksu zakresowego (rozdział 7). Indeks BRI przeznaczony jest do przyspieszania zapytań obliczających wielowymiarowe agregaty zakresowe zdefiniowane w ciągłej przestrzeni cech, wykorzystując do tego celu kodowanie binarne raz operacje na wektorach i macierzach [41, 44];

7. **Podstawowe zadanie H**

Algorytm stronicowania pamięci dla strumieniowych hurtowni danych, wykorzystujący podejście adaptacyjne, bazujące na optymalizacji wielokryterialnej (rozdział 8). Algorytm (w trzech wariantach) został przedstawiony jako główny składnik silnika CUBIT, odpowiedzialny za efektywny transfer danych pomiędzy baza danych a odbiorca; celem algorytmu jest minimalizacja dwóch przeciwstawnych metryk: obciążenia bazy danych oraz czasu oczekiwania na nowe dane. Algorytm wraz z wynikami badan weryfikacyjnych i eksperymentalnych został opisany w [68];

8. **Podstawowe zadanie I**

Metryki jakości usług dla strumieniowych hurtowni danych (rozdział 9). Metryki jakości usług opisują w formie wartości procentowych obciążenie bazy danych (połączenie granulacji zapytań i czasu pomiędzy zadaniami) oraz czas oczekiwania na nowe dane przez użytkownika (połączenie opóźnienia z płynnością w dostawie) – metryki zostały użyte podczas badan eksperymentalnych na algorytmach stronicowania pamięci i opisane w artykule naukowym [68].

Do rozprawy dołączono bardzo obszerny spis literatury, zawierający 153 pozycje naukowe. Wymienione pozycje literaturowe opisują pełny stan techniki reprezentowanej przez literaturę światową.

Doktorant jest współautorem 5 opublikowanych prac, za które uzyskał 410 punktów MNiSW (kolejna nowa publikacja jest aktualnie w procesie recenzowania) oraz jest współautorem pięciu zgłoszeń patentowych w Urzędzie Patentowym Rzeczpospolitej Polskiej.

2.3. Kierunki dalszego rozwoju badań naukowych

W rozprawie doktorskiej opisano model architektury i zdefiniowano założenia dla nowej klasy zaawansowanych hurtowni danych - strumieniowej hurtowni danych kontekstowych. Otwiera to wiele nowych możliwości dalszego ich rozwoju oraz nowych zastosowań praktycznych.

Doktorant precyzyjnie wskazał dalsze plany badawcze do realizacji w przyszłości, nakreślił 5 kierunków dalszego rozwoju badań naukowych:

1. Kontynuacja badan nad strumieniowa adaptacja kostki danych (silnik CUBIT) a w szczególności analiza mechanizmu agregacji w wielowymiarowych oknach zdefiniowanych na wartościach ciągłych. Zaproponowany kierunek

rozwoju implikuje dwa kolejne, które związane są z efektywnym wyznaczaniem oraz dostarczaniem agregatów wielowymiarowych.

2. Rozbudowa koncepcji indeksu BRI (wielowymiarowego bitowego indeksu zakresowego). W rozprawie nakreślono jedynie jego teoretyczne podstawy, które dodatkowo zostały zastrzeżone w postaci patentów. Niemniej, w tym zakresie istnieje wciąż wiele obszarów, które wymagają przeprowadzenia dogłębnych badan, np. struktur indeksu, próba implementacji w środowisku równoległym.

3. Kontynuacja badan nad algorytmami stronicowania pamięci oraz samym mechanizmem stronicowania, którego zadaniem jest zapewnienie efektywnego dostarczania wyników zapytań w formie kostki danych. Zdefiniowano ograniczenia (rozdział 9) zaproponowanych algorytmów i wskazano na konieczność spojrzenia na całość pracujących jednocześnie instancji silnika CUBIT. Zagadnienia te stanowią nowy problem optymalizacyjny.

4. Zastosowaniem koncepcji strumieniowej hurtowni danych kontekstowych do innego problemu rzeczywistego niż omówiony w rozprawie problem wykrywania wycieków paliwa (np. system notowań giełdowych, system monitorowania zmian klimatu, system analizujący dane meteorologiczne).

5. Integracja zaproponowanego modelu strumieniowej hurtowni danych kontekstowych z modelem Spichlerza Agregatów. Koncepcja Spichlerza Agregatów polega na integracji wielu heterogenicznych baz i hurtowni danych w jeden spójny system. Poprzez dopasowanie odpowiednich metod analizy do poszczególnych typów danych, Spichlerz Agregatów integruje dane z wielu różnorodnych hurtowni danych, podobnie jak hurtownia danych integruje dane z wielu różnorodnych baz danych.

Nakreślone w rozprawie metody i modele stanowią istotny wkład w dziedzinie zaawansowanych hurtowni danych, otwierając przy tym szeroki wachlarz perspektyw i możliwości rozwoju.

3. Wskazania – uwagi i pytania

Rozprawa doktorska przygotowana została niezwykle starannie i to zarówno pod względem edytorskim jak i graficznym, co wystawia znakomite świadectwo o dobrym opanowaniu techniki pisania rozpraw naukowych przez doktoranta. Dodatkowo przedstawione analizy, choć z dużą ilością szczegółów, są opisane znakomitym językiem naukowym.

Większość przedstawionych w rozprawie doktorskiej wyników badań została wcześniej opublikowana w formie artykułów naukowych i zgłoszonych patentów. Ten fakt dodatkowo podkreśla wysoką jakość, częściowo już ocenionych, otrzymanych wyników badać naukowych i zaprezentowanych w rozprawie doktorskiej. Dodatkowo należy podkreślić fakt, że w rozprawie są to już wielokrotnie przeanalizowane, zweryfikowane i zredagowane wyniki badań naukowych.

Uwagi

Zwyczajowo recenzenci w swych recenzjach starają się wytknąć drobne niezręczności. Po lekturze rozprawy, zauważyłem w tekście drobne usterki, nie mające jednak istotnego wpływu na jakość naukową rozprawy, np.:

a) do rozprawy nie załączono wykazu używanych skrótów,
b) brak wykazu używanych oznaczeń wykorzystywanych w rozprawie, na str. 32 w tabeli 3.1 są podane „Objaśnienia symboli używanych w notacji matematycznej". Te objaśnienia są ukryte wewnątrz podrozdziału i trudniej je odszukać w tekście pracy.
c) brak centralnego skorowidza podstawowych terminów,
d) — str. 51 (pierwszy wiersz pierwszego akapitu) w zdaniu: „ ... celu niezbędne jest stworzenie modelu przepływu paliwa." użyto żargonowego sformułowania „stworzony", jednak zręczniej byłoby użyć jednej z form, np. model „zbudowany", „zaprojektowany", „przedstawiony", „opisany", itp.
— str. 104, drugi akapit „ ...stworzono specjalne środowisko badawcze", zręczniej byłoby użyć jednej z form „zbudowano", „zrealizowano" w miejsce „stworzono".

Pytania do doktoranta

Doktorant w podsumowaniu wyników rozprawy wskazał na otwarte problemy badawcze, np. na możliwość adaptacji zaproponowanych rozwiązań (m. in. zdefiniowanych modeli i metod przetwarzana strumieni danych w hurtowniach danych oraz aspekty ich stosowania w systemach wykrywających anomalie) do innych problemów świata rzeczywistego. W zakończeniu rozprawy czytamy:

„Głównym problemem badawczym, a zarazem bezpośrednią motywacją do podjęcia tematyki rozprawy, było wykrywanie wycieków z podziemnych zbiorników paliw. Dokładne poznanie charakterystyki zjawisk zachodzących na stacjach paliw pozwoliło na identyfikacje podstawowych anomalii rzeczywistych (zjawisk niepożądanych) oraz na dokonanie ich klasyfikacji pod kątem lokalności i istotności.

Weryfikacje tez rozprawy polegały na budowie modeli i analizie powiązanych zagadnień teoretycznych. Sposób empiryczny opierał się na wynikach eksperymentów oraz wnioskach sformułowanych przy pracy z rzeczywistymi obiektami przemysłowymi. Punktem wyjścia do obu podejść był przykład motywujący - system dystrybucji i składowania paliw płynnych."

1. Czy zdefiniowane w rozprawie modele i metody przetwarzana strumieni danych w hurtowniach danych oraz aspekty ich stosowania w systemach wykrywających anomalie będzie można bezpośrednio wykorzystać do innych systemów?
2. W jakim stopniu zaproponowane w rozprawie rozwiązania są uniwersalne? Dowód trzech tez postawionych w rozprawie był oparty tylko na jednym przykładzie systemu.

3. Czy zdefiniowanie charakterystyk zjawisk zachodzących na stacjach paliw pozwoliłoby na identyfikacje podstawowych zjawisk niepożądanych i na dokonanie ich klasyfikacji w innych systemach rzeczywistych?
4. Czy można wykorzystać koncepcję strumieniowej hurtowni danych kontekstowych do innych problemów rzeczywistych niż omówiony w rozprawie problem wykrywania wycieków paliwa (np. system notowań giełdowych, system monitorowania zmian klimatu, system analizujący dane meteorologiczne)?

4. Podsumowanie – końcowe wnioski

Stwierdzam, że rozprawa doktorska mgr. inż. Krzysztofa Pasteraka zawiera istotne elementy nowości naukowe w dyscyplinie informatyka techniczna i telekomunikacja. Problemy badawcze rozważane w rozprawie związane z detekcją anomalii krytycznych z zastosowaniem modeli i metod strumieniowych hurtowni danych mają charakter naukowy i mają duże znaczenie praktyczne, a uzyskane wyniki stanowią znaczący wkład w rozwój zaawansowanych hurtowni danych.

Doktorant w sposób profesjonalny postawił problemy badawcze, wykazał się biegłością w zakresie prezentowanej tematyki przedmiotu i doskonałością warsztatu badawczego. Przeprowadzona analiza otrzymanych wyników i wzorowa prezentacja, dowodzi tym samym dojrzałości naukowej doktoranta. Cel rozprawy doktorskiej został osiągnięty, a postawione tezy naukowe zostały udowodnione. Dodatkowo - staranność, z jaką została przygotowana rozprawa wystawia jednoznacznie pozytywne świadectwo dla Autora.

Z całkowitym przekonaniem uważam, że praca doktorska mgr. inż. Krzysztofa Pasteraka z dużym naddatkiem spełnia wszelkie warunki formalne stawiane przez obowiązującą ustawę o stopniach i tytułach naukowych (art. 13 ustawy z dnia 14 marca 2003) i moja ocena rozprawy pod względem trzech podstawowych kryteriów jest następująca:

A. Czy rozprawa zawiera oryginalne rozwiązanie problem naukowego?

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

B. Czy po przeczytaniu rozprawy zgadzasz się, że kandydat posiada ogólną wiedzę teoretyczną w dyscyplinie Informatyka techniczna i telekomunikacja?

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

C. Czy kandydat umiejętność samodzielnego prowadzenia pracy naukowej?

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

i w związku z tym **wnioskuję o dopuszczenie doktoranta do dalszych etapów przewodu doktorskiego.**

Ze względu na bardzo wysoką wartość merytoryczną rozprawy, dojrzałość naukową niezwykle pracowitego doktoranta i jego dorobek naukowy (rozprawa doktorska, 5 publikacji[1] z nadanymi numerami DOI o łącznej liczbie 410 ministerialnych punktów i 5 zgłoszeń patentowych[2]), rekomenduję wyróżnienie i wnoszę również do Rady Dyscypliny o rozważenie możliwości **wyróżnienia rozprawy doktorskiej.**

Zygmunt Mazur

Załącznik: Bibliometria doktoranta

[1] Lista publikacji w załaczniku do recenzji

[2] Lista zgłoszonych patentów do Urzędu Patentowego Rzeczpospolitej Polskiej w załaczniku do recenzji

Załącznik do recenzji

Dorobek naukowy mgr inż. Krzysztofa Pasteraka

- Publikacje 5
- Patenty 5

Bibliometria

- h-index (Cytowania Scopus) 2 - h-index (Cytowania Scopus)
- h-index (Cytowania WoS) 2 - h-index (Cytowania WoS)
- Sumaryczny IF - 3,768
- Sumaryczny SNIP - 2,534
- Sumaryczny CiteScore - 9,5
- **Sumaryczna punktacja MNiSW - 410**

Artykuły z czasopism

2017 The TUBE algorithm: discovering trends in time series for the early detection of fuel leaks from underground storage tanks, Gorawski Marcin, Gorawska Anna, Pasterak Krzysztof, Expert Systems with Applications, 2017, vol. 90, s.356-373. DOI:10.1016/j.eswa.2017.08.016

Rozdziały z monografii

2016 Anomaly detection in data streams: the petrol station simulator , Gorawska Anna, Pasterak Krzysztof, W: Beyond databases, architectures and structures : Advanced technologies for data mining and knowledge discovery. 12th International conference, BDAS 2016, Ustroń, Poland, May 31 - June 3, 2016. Proceedings / Kozielski Stanisław [i in.] (red.), 2016, Springer, s.727-736, ISBN 978-3-319-34098-2. DOI:10.1007/978-3-319-34099-9_57

2015 Liquefied petroleum storage and distribution problems and research thesis , Gorawski Marcin, Gorawska Anna, Pasterak Krzysztof, W: Beyond databases, architectures and structures : 11th International conference. BDAS 2015, Ustroń, Poland, May 26-29, 2015. Proceedings, Communications in Computer and Information Science, 2015, vol. 521, Springer, s.540-550, ISBN 978-3-319-18421-0. DOI:10.1007/978-3-319-18422-7_48

2015 Research and analysis of the stream materialized aggregate list , Gorawski Marcin, Pasterak Krzysztof, W: Computer science and its applications : 5th IFIP, TC 5 International Conference CIIA 2015, Saida, Algeria, May 20-21, 2015. Proceedings / Abdelmalek A. [i in.] (red.), IFIP Advances in Information and Communication Technology, 2015, vol. 456, Springer, s.269-278, ISBN 978-3-319-19577-3. DOI:10.1007/978-3-319-19578-0_22

2014 A survey of data stream processing tools , Gorawski Marcin, Gorawska Anna, Pasterak Krzysztof, W: Information sciences and systems 2014 : Proceedings of the 29th International Symposium on Computer and Information Sciences, 2014, Springer, s.295-303, ISBN 978-3-319-09464-9. DOI:10.1007/978-3-319-09465-6_31

Patenty – wynalazki

- Układ do kalibracji stanu zbiornika paliw płynnych , Gorawski Marcin, Skrzewski Mirosław, Gorawska Anna, Pasterak Krzysztof, Wynalazek, Chroniony, Numer zgłoszenia (w pierwszym kraju zgłoszenia powyżej): 421 004, Numer patentu/prawa: 237 471, Data zgłoszenia (w pierwszym kraju zgłoszenia powyżej): 27-03-2017, Data udzielenia prawa: 19-04-2021

- Sposób wykrywania niezgodności w procesie dostaw paliw płynnych na stacjach paliw , Gorawski Marcin, Skrzewski Mirosław, Gorawska Anna, Pasterak Krzysztof, Gorawski M., Wynalazek, Chroniony, Numer zgłoszenia (w pierwszym kraju zgłoszenia powyżej): 421 877, Numer patentu/prawa: 235 154, Data zgłoszenia (w pierwszym kraju zgłoszenia powyżej): 12-06-2017, Data udzielenia prawa: 14-02-2020, Publikacja patentu/wzoru: [WUP 01-06-2020]

- Sposób inteligentnego wykrywania anomalii w torze przepływu paliwa na stacjach paliw , Gorawski Marcin, Gorawska Anna, Pasterak Krzysztof, Wynalazek, Chroniony, Numer zgłoszenia (w pierwszym kraju zgłoszenia powyżej): 421 264, Numer patentu/prawa: 235 055, Data zgłoszenia (w pierwszym kraju zgłoszenia powyżej): 10-04-2017, Data udzielenia prawa: 07-01-2020, Publikacja patentu/wzoru: [WUP 18-05-2020]

- Sposób ekstrakcji i transformacji strumieniowych danych pomiarowych wykorzystujący obliczenia równoległe , Gorawski Marcin, Gorawska Anna, Pasterak Krzysztof, Gorawski M., Wynalazek, Chroniony, Numer zgłoszenia (w pierwszym kraju zgłoszenia powyżej): 423 219, Numer patentu/prawa: 233 157, Data zgłoszenia (w pierwszym kraju zgłoszenia powyżej): 20-10-2017, Data udzielenia prawa: 06-05-2019, Publikacja patentu/wzoru: [WUP 30-09-2019]

- Sposób magazynowania i agregowania wielowymiarowych strumieniowych danych pomiarowych z uwzględnieniem aspektu lokalności , Gorawski Marcin, Gorawski M., Gorawska Anna, Pasterak Krzysztof, Wynalazek, Chroniony, Numer zgłoszenia (w pierwszym kraju zgłoszenia powyżej): 421 668, Numer patentu/prawa: 232 115, Data zgłoszenia (w pierwszym kraju zgłoszenia powyżej): 22-05-2017, Data udzielenia prawa: 06-12-2018, Publikacja patentu/wzoru: [WUP 31-05-2019]