POLITECHNIKA POZNAŃSKA

WYDZIAŁ INFORMATYKI I TELEKOMUNIKACJI
(FACULTY of COMPUTING AND TELECOMMUNICATIONS)
ul. Piotrowo 3, 60-965 Poznań, tel. +48 61 665 3420, fax +48 61 665 3421
e-mail: office_dcatf@put.poznan.pl, www.fcat.put.poznan.pl

Recenzja rozprawy doktorskiej mgra Macieja Brzeskiego
pt. Analiza podobieństwa procesów ETL i pochodzenia danych w kontekście zarządzania hurtownią danych

Wykonał: **dr hab. inż. Robert Wrembel, prof. PP**

Politechnika Poznańska
Wydział Informatyki i Telekomunikacji
Instytut Informatyki
Piotrowo 2
60-965 Poznań
tel: +48 601 392346, +48 453 031788, +48 61 665 2991, +48 665 2999
e-mail: robert.wrembel@cs.put.poznan.pl

1. Tematyka i zarys problemu

Integracja danych (ID) jest jednym z ważniejszych tematów badawczych. Celem technik ID jest udostępnianie danych o różnych strukturach i modelach, pochodzących z wielu źródeł danych (ŹD). Techniki te czyszczą i przekształcają dane do ujednoliconego formatu, odpowiedniego dla zaawansowanych analiz. Prace badawczo-rozwojowe zaowocowały kilkoma standardowymi architekturami ID. We wszystkich z nich, dane są przenoszone ze ŹD do systemu zintegrowanego za pomocą warstwy integracyjnej, która uruchamia procesy ID. Są to złożone przepływy pracy (workflows) składające się z wielu zadań. Zadania te odpowiadają za pobieranie danych ze ŹD i udostępnienie ich aplikacjom analitycznym.

Ponieważ procesy ID są podstawowymi komponentami wszystkich architektur integracji danych, metody ich projektowania są badane i rozwijane od dziesięcioleci. Wiele z tych metod zostało zaimplementowanych w komercyjnym i otwartym oprogramowaniu do ID. Mimo znacznej liczby rozwiązań w tym zakresie, projektowanie i zarządzanie procesami ID jest nadal trudne i czasochłonne.

Pomimo prac naukowych prowadzonych od dziesięcioleci przez wiele uznanych na świecie ośrodków, nadal istnieją otwarte problemy badawcze i technologiczne, do których zalicza się m.in.: (1) optymalizację procesów ID, (2) zarządzanie funkcjami zdefiniowanymi przez użytkownika (UDF) w procesach ID, (3) zarządzanie ewolucją procesów ID na skutek ewolucji źródeł danych, (4) zarządzanie historią przetwarzania danych (data lineage / data provenance) w procesach ID.

Ze wspomnianych ww. otwartych problemów, niniejsza rozprawa adresuje problem 1 i 4.

Optymalizacja procesów ID ma fundamentalny wpływ na wydajność całego systemu informatycznego. Problem jest trudny i pomimo dziesiątek rozwiązań opublikowanych w najlepszych czasopismach i najlepszych konferencjach nie rozwiązano go w pełni.

Data lineage to zestaw technik dokumentujących pełen cykl życia danych. Techniki te są kluczowe dla polityk zarządzania danymi (data governance) i są wymagane przepisami regulującymi rynki finansowe. W literaturze naukowej dotychczas zaproponowano wiele koncepcji budowania i utrzymywania data lineage. W ostatnim czasie, naukowcy proponują wykorzystanie algorytmów grafowych do znajdowania i reprezentowania data lineage. Jednak ani starsze, ani najnowsze koncepcje nie rozwiązują do końca problemu.

Wspomniane rozwiązania w zakresie optymalizacji procesów ID i lineage są publikowane na najlepszych na świecie konferencjach, np. VLDB, ICDE, SIGMOD i w najlepszych czasopismach, np. VLDB Journal, ACM Journal of Data and Information Quality, Data & Knowledge Management, IEEE Transactions on Knowledge and Data Engineering, co wskazuje na wagę tej problematyki.

W tym kontekście, tematykę recenzowanej rozprawy doktorskiej oceniam jako bardzo ważną z naukowego i praktycznego punktu widzenia. Rozwiązywane w rozprawie problemy są właściwe dla rozpraw doktorskich.

2. Ocena osiągnięć w świetle Ustawy

Moją ocenę osiągnięć Doktoranta opieram o:

- *Art. 186 Ustawy* z dnia 20 lipca 2018 r. – *Prawo o szkolnictwie wyższym i nauce*, który stwierdza w P.1.: "Stopień doktora nadaje się osobie, która ... 3) posiada w dorobku co najmniej: a) 1 artykuł naukowy opublikowany w czasopiśmie naukowym lub w recenzowanych materiałach z konferencji międzynarodowej, które w roku opublikowania artykułu w ostatecznej formie były ujęte w wykazie sporządzonym zgodnie z przepisami wydanymi na podstawie art. 267 ust. 2 pkt 2 lit. b ...";
- *Art. 187 Ustawy*: "1. Rozprawa doktorska prezentuje ogólną wiedzę teoretyczną kandydata w dyscyplinie albo dyscyplinach oraz umiejętność samodzielnego prowadzenia pracy naukowej lub artystycznej. 2. Przedmiotem rozprawy doktorskiej jest oryginalne rozwiązanie problemu naukowego, oryginalne rozwiązanie w zakresie zastosowania wyników własnych badań naukowych w sferze gospodarczej lub społecznej albo oryginalne dokonanie artystyczne. 3. Rozprawę doktorską może stanowić praca pisemna, w tym monografia naukowa, zbiór opublikowanych i powiązanych tematycznie artykułów naukowych ...".

W tym kontekście, przedstawiona **rozprawa doktorska spełnia wymagania** zapisane w Art. 186 Ustawy, ponieważ Doktorant posiada w dorobku dwa artykuły opublikowane w miejscach wskazanych zapisem ustawy, tj.:

- [1] M. Brzeski, A. Roman: Measuring Similarity Between ETL Processes Using Graph Edit Distance. Schedae Informaticae 32, 2023,
- [2] M. Brzeski, A. Roman: Inferring Missing Data Lineage Links from Schema Metadata Using Transformer-Based Models. Proc. of the VLDB Endowment, 2025.

Spełnione są także wszystkie wymagania Art. 187 Ustawy.

Pomimo tego, w mojej **ocenie** osiągnięć Doktoranta stwierdzam, że wymagania Art. 187 są spełnione w stopniu minimalnym. Artykuł [1] został opublikowany w czasopiśmie za 20 pkt., a artykuł [2], mimo że jest "warty" 200 pkt. został opublikowany na workshop'ie 6th Applied AI for Database Systems and Applications (AIDB).

3. Struktura rozprawy

Rozprawa mgra Macieja Brzeskiego jest obszerna (232 str.). Rozdziały 1-4 i 6 prezentują podstawy teoretyczne i technologiczne, na których bazują kontrybucje rozprawy. Rozdział 5 przedstawia wyniki badawcze Doktoranta w zakresie optymalizacji procesów ID, a rozdział 7 - wyniki dotyczące rozwiązań data lineage. Rozdział 8 podsumowuje rozprawę.

Pomimo, że rozprawa jest obszerna i zawiera bardzo szczegółowe informacje wstępne, uważam, że nie zachowano w niej równowagi pomiędzy częścią wstępno-teoretyczną - pięć rozdziałów o łącznej objętości 119 stron, a rozdziałami prezentującymi kontrybucję rozprawy - dwa rozdziały o łącznej objętości 63 stron.

W rozprawie zaadresowano trzy główne obszary badawcze, tj.: (1) projektowanie i optymalizowanie procesów ID (w rozprawie nazywanych procesami ETL), (2) data lineage i (3) podobieństwo grafów. W każdym z tych obszarów istnieje mnogość rozwiązań. Niestety, przegląd literatury dotyczący tych obszarów zawarty w rozprawie jest zdecydowanie zbyt wąski. Brakuje tu np.:

- prac dot. wydajności procesów ETL, np.:
  - M. Gorawski, M. Gorawski, S. Dyduch: *Use of Grammars and Machine Learning in ETL Systems That Control Load Balancing Process*. Int. Conference on High Performance Computing and Communications & IEEE Int. Conference on Embedded and Ubiquitous Computing (HPCC/EUCHPCC/EUC), 2013,
  - N. Kumar, P.S. Kumar: An Efficient Heuristic for Logical Optimization of ETL Workflows. In VLDB Workshop on Enabling Real-Time Business Intelligence, 2010
  - A. Simitsis, P. Vassiliadis, T. Sellis: Optimizing ETL Processes in Data Warehouses. Int. Conference on Data Engineering (ICDE), 2005
- prac dot. projektowania procesów ETL, np.:
  - J. Awiti, A.A. Vaisman, E. Zimányi: *Design and implementation of ETL processes using BPMN and relational algebra*. Data & Knowledge Engineering 129, 2020,

- J. Awiti, A.A. Vaisman, E. Zimányi: From Conceptual to Logical ETL Design Using BPMN and Relational Algebra. Int. Conf. Big Data Analytics and Knowledge Discovery (DAWAK), 2019,
- P. Jovanovic, O. Romero, A. Simitsis, A. Abelló: Incremental Consolidation of Data-Intensive Multi-Flows. IEEE Transactions on Knowledge and Data Engineering 28, 5, 2016,

- podobieństwa procesów ETL, np.:
  - A. Wojciechowski: *ETL workflow reparation by means of case-based reasoning*. Information Systems Frontiers 20(1), 2018,
  - L. Muñoz, J.-N. Mazón, J. Trujillo: Measures for ETL Processes Models in Data Warehouses. Int. Workshop on Model Driven Service Engineering and Data Quality and Security, 2009

Serwis *dblp.org* udostępnia setki artykułów dot. ww. obszarów badawczych.

4. Osiągnięcia rozprawy

Do najważniejszych osiągnięć rozprawy zaliczam:

- zdefiniowanie metody określania podobieństwa procesów ETL wraz z autorskimi modyfikacjami znanych algorytmów, tj. Hungarian-A*-top, Hungarian-A*-cut, DAG-Edit;
- zaproponowanie metody odkrywania brakujących powiązań data lineage, jednak przy dość silnych ograniczeniach co do struktury obiektów bazy danych (por. uwagi w Sekcji 5.1 niniejszej recenzji).

Każda z powyższych koncepcji została opracowana teoretycznie, zaimplementowana i oceniona eksperymentalnie. Zatem od strony metodyki badań naukowych osiągnięcia rozprawy są poprawne.

5. Uwagi

Poniżej przedstawiam uwagi podzielone na merytoryczne i pozostałe (mniej ważne).

5.1. Uwagi merytoryczne

Nie zgadzam się z definicją hurtowni danych podaną na str. 11: "W tym ujęciu hurtownia danych powstaje w sposób iteracyjny jako zbiór magazynów danych (ang. data marts), zorganizowanych wokół konkretnych obszarów biznesowych, takich jak sprzedaż, czy logistyka". Hurtownia danych jest centralnym repozytorium danych (w jednolitej postaci, oczyszczonych, zunifikowanych), na którym to repozytorium są budowane data marts, czyli

tzw. tematyczne hurtownie danych. Data marts zawierają wyłącznie dane zagregowane dot. podzbioru danych z hurtowni danych.

Strona 16: nie zgadzam się z definicją ELT: "... wariant ELT, w którym dane najpierw są ładowane do hurtowni danych ...". W architekturze ELT dane są wczytywane do tzw. operacyjnej składnicy danych (operational data store, data stage, staging area, landing pad), a z niej procesy TL transformują dane i wczytują do hurtowni. Składnica danych i hurtownia danych mogą fizycznie być implementowane w tym samym repozytorium danych - tym samym systemie bazy danych, ale logicznie są to odrębne przestrzenie danych.

Sekcja 2.3.2: opisane tu podejście zakłada, że pochodzenie danych (data lineage) jest konstruowane w czasie budowania procesów ID. Takie założenie jest często nierealne ponieważ zwykle nie mamy dostępu do procesów ID, a istnieje konieczność odtworzenia pochodzenia danych. Jak ma się podejście opisane w Sekcji 2.3.2 do kontrybucji rozprawy? Jeśli zakłada się, że data lineage jest znane, wówczas jaki problem rozwiązuje Doktorant w rozprawie?

Na stronie 111 wprowadzono termin "istotne transformacje" nie podając jednak jego definicji. Termin ten wydaje się być dość ważny w kontekście określania podobieństwa grafów. W jaki sposób zmierzyć "istotność transformacji"?

Sekcja 5.4 zawiera kontrybucję rozprawy, jednak jej zawartość jest ogólnikowa. Wykorzystano tu znany algorytm DBSCAN do grupowania. W jakiej postaci proces ETL (graf) jest przekazywany do algorytmu? Ile wymiarów ma przestrzeń, w której dokonuje się grupowania?

Sekcja 5.4.2: jak jest konstruowany wektor reprezentujący proces? Czy jego długość jest stała dla wszystkich procesów? Wydaje się, że nie powinna być stała ponieważ procesy ETL mają różną złożoność. Jeśli jednak zakłada się stałą długość tych wektorów, to w jaki sposób są one normalizowane? Czy wyzwaniem są procesy złożone z tysięcy zadań? Także i w tej sekcji opis jest zbyt ogólnikowy, aby można go zrozumieć.

Sekcja 5.6.5: autor stwierdza, że kontrybucją rozprawy są algorytmy: Hungarian-A*-top, Hungarian-A*-cut i DAG-Edit, ale w rozprawie nie zostały one opisane ani zilustrowane przykładami wykorzystania. W sekcji 5.6.4 pojawiają się natomiast wyniki oceniające te algorytmy.

Metodyka oceny eksperymentalnej: moim zdaniem trudno ocenić zaproponowane algorytmy ponieważ nie jest znana wartość prawdziwa (ground truth). Czy ona istnieje, a jeśli tak, to w jaki sposób jest wyznaczana? Brak wskazania ground truth i procedury konstruowania tej bazy

odniesienia uniemożliwia ocenienie jakości wyników produkowanych przez zaproponowane rozwiązania.

Sekcja 7.1.3, stwierdzenie: "W tej pracy koncentrujemy się na transformacjach przetwarzających jedno źródło (ang. single-source calculation), w których każda kolumna docelowa jest pochodną dokładnie jednej kolumny źródłowej" - moim zdaniem jest to zbyt silne ograniczenie, wyłącza bowiem z analizy klasyczne i bardzo częste scenariusze transformacji danych, w których z jednej wartości atrybutu powstaje wiele wartości, np. rozbicie adresu na jego składowe.

Strona 153, stwierdzenie: "Jednocześnie wiele niepoprawnych kandydatów wykazuje powierzchniowe podobieństwo wynikające z konwencji nazewniczych lub skrótów, co czyni je pozornie prawdopodobnymi". Z tego powodu, nie można jedynie polegać na nazewnictwie obiektów ani na metadanych, a w Sekcji 7.1.5 założono, że "modele działają wyłącznie na metadanych schematów". Stwierdzenia te i założenia są przeciwstawne i wymagają szczegółowego wyjaśnienia.

Sekcja 7.3.2, stwierdzenie: "Ostatecznie zdecydowałem się na ujednolicenie nazw kolumn, sprowadzając je do małych liter i dzieląc na osobne wyrazy" - w jaki sposób nazwy dzielono na wyrazy? W wielu przypadkach nie da się dokonać takiego podziału automatycznie - do tego potrzebna jest semantyka nazewnictwa, bo na jej podstawie w pewnych przypadkach da się określić granice podziałów. Przykładowo, skąd algorytm dzielący "wie" jak podzielić zlepek znaków "fechaDenaCimiento"?

W Rozdziale 7 brakuje opisu scenariuszy w jakich zaproponowane rozwiązanie jest w stanie wykryć zerwane połączenia. Brak takich scenariuszy wpływa szczególnie źle na czytelność Sekcji 7.5, ponieważ nie wiadomo co jest testowane.

Opis w Sekcji 7.5.1 sugeruje mi, że chodzi o tzw. schema matching, a nie o data lineage. Schema matching to techniki umożliwiające identyfikowanie odpowiadających sobie obiektów w różnych schematach baz danych.

W całym Rozdziale 7 brakuje także opisu co jest źródłem prawdy - skąd wiadomo, że odkryte scenariusze odpowiadają rzeczywistości? Brakuje także porównania z istniejącymi rozwiązaniami. Skąd wzięto ground truth dla tysięcy schematów? Jak przygotowano dane uczące na podstawie tysięcy schematów?

Przedstawiona rozprawa doktorska jest typu wdrożeniowego. Niestety, w rozprawie pominięto opis wdrożenia rozwiązania. Doktorant wspomina jedynie o charakterze wdrożeniowym w Rozdziale 1, o wdrożeniu - jedno zdanie w Sekcji 8.1, a w Sekcji 8.2 prezentuje niestety dość oczywistą problematykę wdrożeniową, bez szczegółowego

omówienia wdrożenia zaproponowanego rozwiązania. Na tej podstawie trudno ocenić wartość rozprawy w kontekście wdrożonego rozwiązania.

5.2. Pozostałe uwagi

Strona 7: wprowadzono skrót DAG bez jego uprzedniego wyjaśnienia.

Strona 13: pozostawiony tekst "A wcześniej pisałeś".

Strona 15: nie zdefiniowano terminu "silnik przetwarzania".

Strona 17 i dalej w całej rozprawie: "informacje" to nie to samo co "dane". W hurtownach danych, bazach danych i innych repozytoriach danych przechowujemy dane.

Strona 18: sformułowanie "job ETL" jest niewłaściwe - w j. polskim nazywamy je zadaniami ETL.

Sekcja 2.2.2: do metadanych technicznych zaliczamy także te, które sterują kontrolą dostępu.

Strona 39, operacja "Wstawienie wierzchołka", pierwsza pozycja listy wypunktowanej: symbol T chyba powinien być zastąpiony przez 'duże T pisane'.

Błędy składu tekstu: np. str. 71, 142, 163 - różne odległości między akapitami.

Strona 123, ostatnia linia: o jakie pary pól chodzi?

Wzór 5.16: brakuje przykładu ilustrującego.

Sekcja 5.5.3: brakuje przykładów.

Strona 127: "Algorytmy bazowe" i "Propozycje własne" są tytułami, więc nie stosujemy kropek na końcach.

Sekcja 6.3.2: brak angielskiego odpowiednika "mechanizmu uwagi" - we wcześniejszych rozdziałach pracy takie odpowiedniki były podawane.

Strona 152: co to jest "system danych"?

Tytuł sekcji 7.3.1 "Organizacja danych" jest mylący - sugeruje fizyczną organizację danych, np. indeksowanie, partycjonowanie, optymalizację rozmieszczenia segmentów danych na dysku.

Brakuje wyraźnego zaznaczenia rozdziałów wnoszących kontrybucję rozprawy.

POLITECHNIKA POZNAŃSKA

WYDZIAŁ INFORMATYKI I TELEKOMUNIKACJI
(FACULTY of COMPUTING AND TELECOMMUNICATIONS)
ul. Piotrowo 3, 60-965 Poznań, tel. +48 61 665 3420, fax +48 61 665 3421
e-mail: office_dcatf@put.poznan.pl, www.fcat.put.poznan.pl

6. Ocena końcowa

Podsumowując, rozprawa doktorska mgra Macieja Brzeskiego porusza bardzo ciekawą problematykę, która znajduje się w obszarze zainteresowań nie tylko czołowych ośrodków naukowych na świecie, ale także przemysłu oprogramowania i biznesu. Niestety, praca bardzo ogólnie prezentuje wyniki prac badawczych Doktoranta, zawarte w Rozdziałach 5 i 7. Dorobek publikacyjny Doktoranta jest skromny.

W ogólności, sposób formułowania problemów badawczych i metodyka ich rozwiązania, zawarte w recenzowanej rozprawie, są właściwe dla badań naukowych w informatyce i rozpraw doktorskich. Zastosowana metodyka bazuje na analizie teoretycznej problemu, budowie modelu matematycznego, implementacji rozwiązania i jego ocenie eksperymentalnej.

Stosowalność rozwiązań opracowanych w ramach rozprawy została uwiarygodniona dwoma publikacjami naukowymi, tym samym spełniony został warunek *Art. 186 Ustawy* z dnia 20 lipca 2018 r. – *Prawo o szkolnictwie wyższym i nauce*. Powtarzam tu moją opinię, że dorobek publikacyjny Doktoranta jest skromny. Cel rozprawy został osiągnięty, choć w minimalnym zakresie.

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

Signed by / Podpisano przez:

Robert Andrzej Wrembel
Politechnika Poznańska

Date / Data: 2026-03-29 12:25