Politechnika Wrocławska

Katedra Sztucznej Inteligencji

Wrocław, 13 maja 2026r.

Prof. dr hab. inż. Przemysław Kazienko
Katedra Sztucznej Inteligencji
Wydział Informatyki i Telekomunikacji
Politechnika Wrocławska
Email: kazienko@pwr.edu.pl
WWW: http://kazienko.eu

RECENZJA

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

Rozprawę napisano pod kierunkiem prof. dr hab. Adama Romana i mgra Dawida Dudy (promotor pomocniczy) na Uniwersytecie Jagiellońskim i złożono w 2025r.

Recenzję wykonano na zlecenie Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Uniwersytetu Jagiellońskiego na podstawie decyzji Rady z dnia 22 stycznia 2026r. i dotyczy ona procesu o nadanie stopnia doktora w dziedzinie nauk inżynieryjno-technicznych, w dyscyplinie informatyka techniczna i telekomunikacja.

I. Przedmiot, problematyka i charakter rozprawy

Tematyka rozprawy mieści się w szerokiej, rozwijanej od kilkudziesięciu lat dziedzinie hurtowni danych i dotyczy dwóch, wybranych problemów związanych z procesami ETL (*Extract, Transform, Load*) czyli przygotowania i przetwarzania danych, np. dla potrzeb hurtowni, ale nie tylko dla nich. Owe problemy to: (1) pochodzenie danych (*data lineage*), w tym także wnioskowanie o owym pochodzeniu w przypadku braku odpowiedniej informacji oraz (2) analiza podobieństwa procesów ETL, co łącznie ma na celu zwiększenie wiedzy i zaufania do danych zgromadzonych w hurtowni. Tematykę można określić jako niszową, ale bez wątpienia, zawiera się w dyscyplinie informatyki, w której pracę złożono. W ramach pierwszego problemu Autor ograniczył się do binarnego wnioskowania do zależności między parami elementów składowych struktur danych. Drugie zagadnienie zostało przedefiniowane jako problem porównywania drzew i grafów.

Rozważana problematyka, tj. podobieństwo drzew i grafów oraz identyfikacja odpowiadających sobie opisów tekstowych – metadanych dla struktur danych, nie jest nowa, aczkolwiek ich zastosowanie do procesów ETL jest wartościowym wkładem Autora. Owo odniesienie do znanych i relatywnie starych problemów w większości dotyczy także metod stosowanych do ich rozwiązania. W efekcie można określić rozwijaną tematykę jako do pewnego stopnia niszową i poza głównym nurtem badań na świecie. W ramach tego wnioskowanie o powiazaniach

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Katedra Sztucznej Inteligencji
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
T: +48 71 320 24 54
www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl
REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434

struktur danych na podstawie analizy tekstowych metadanych z wykorzystaniem uczenia maszynowego jest najbardziej aktualne.

Doktorat realizowano jako wdrożeniowy, we współpracy z firmą *Informatica*. Dzięki temu praca ma inspiracje praktyczne, tzn. próbuje rozwiązywać rzeczywiste zagadnienia, transformując je do problemów o charakterze bardziej naukowym, czasami wręcz teoretycznym.

II. Hipotezy i cele pracy

Cztery cele pracy zostały opisowo sformułowane w pkt. 1.3.1 (s. 4). Obejmują one:

1. Analizę formalną odległości edycyjnej drzew i grafów, ujednolicenie definicji oraz wykazanie relacji między wybranymi problemami podobieństwa struktur.
2. Opracowanie efektywnej metody wnioskowania o pochodzeniu danych na podstawie struktury schematów baz danych.
3. Opracowanie narzędzia do mierzenia podobieństwa procesów ETL przy użyciu odległości edycyjnej dla drzew i grafów.
4. Implementację i ewaluację narzędzia do wnioskowania brakujących powiązań pochodzenia danych między tabelami i kolumnami.

Należy to tutaj zauważyć, że cele określone jako „*implementacja*" lub „*opracowanie narzędzia*" można by formalnie traktować bardziej jako cele inżynierskie niż badawcze lub naukowe. Z drugiej jednak strony, treść relewantnych rozdziałów odnosi je raczej do „metod" lub „rozwiązań" niż tylko inżynierskich implementacji, co wprost wypełnia założenie ich naukowości.

Formalnie nie sformułowano hipotez, ale osobiście nie uważam, że prace doktorskie i badania naukowe muszą posiadać sformułowane hipotezy i w tym konkretnym przypadku umiarkowanie mi tego brakuje w ocenianej monografii.

III. Oryginalne osiągnięcia

Praca zawiera kilka wartościowych osiągnięć. Ze względu na pewne braki w odesłaniach literaturowych, ich pełna ocena wymagałaby dużej pracy związanej z wyszukiwaniem literatury oraz porównywaniem treści monografii z treściami artykułów innych autorów. Z tego powodu skupiłem się jedynie na deklaracjach Autora dotyczących jego wkładu, wyrażonych w pkt.1.5 (s. 6-7):

1. Rozdział 3. Zadeklarowanym i rzeczywistym osiągnięciem Autora jest pkt. 3.4 i 3.5. Autor twierdzi, że w pkt. 3.4 „*zunifikował różne definicje*". Ogólnie bym się z tym zgodził, pomimo tego, że nie do końca wiadomo, na czym owa unifikacja polegała. Można więc tutaj nieco dyskutować, na ile treści zawarte w 3.4 są oryginalnym wkładem Autora, a na ile są one pewnym opracowaniem wcześniej znanych koncepcji. Przykładowo, wyrównywanie drzew jest w dużej mierze wzięte z [66] z 1995r., co nie

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Katedra Sztucznej Inteligencji
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
T: +48 71 320 24 54
www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl
REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434

zostało wprost zaznaczone. Ogólnie wątpliwości można mieć do wszystkich trzech rozważanych podejść, tj. wyrównywania, ograniczonej odległości edycyjnej i mapowania poprzez naddrzewo. W pracy nie podano odpowiednich wskazań literaturowych, a literatura w tej dziedzinie jest bardzo rozległa i czasami sięga lat 90. ubiegłego wieku. Przykładowo, problem największych wspólnych poddrzew był rozważany m.in. przez Tatsuya Akutsu, czego nie dowiemy się przy definicji 3.30. Ten konkretny problem jest bardzo ważny w kolejnym pkt. 3.5. Generalnie więc, ze względu na braki w odesłaniach literaturowych oraz ze względu na brak dobrze opisanej sekcji analizy literaturowej i wskazania tego, co inni zrobili, a czego nie zrobili (i dlatego Autor chce coś rozwiązać), nieco trudno rozdzielić dokonania Autora od tych wziętych z literatury.

2. Jeżeli chodzi o pkt. 3.5 Autor twierdzi, że podważa twierdzenie z pracy [1], przy czym nie bardzo wiadomo, które twierdzenie (jest tam 31 twierdzeń, lematów i wniosków). Stwierdzenie Autora jest stosunkowo 'mocne' i wartościowe, ale mamy tutaj chyba pewną niespójność: Autor twierdzi, że „*Nie jest możliwe istotne przyspieszenie rozwiązania jednego problemu dzięki znajomości rozwiązania drugiego*" (s. 71) i że „*Wynik ten podważa twierdzenie przedstawione w pracy [1], w której dowodzono, że rozwiązanie jednego z tych problemów (największego wspólnego poddrzewa lub najmniejszego wspólnego naddrzewa) umożliwia bezpośrednie skonstruowanie rozwiązania drugiego*". Jak więc brak przyspieszenia lub to że oba podejścia „*mimo pozornej podobieństwa definicji, prowadzą do zasadniczo różnych rezultatów*" (s. 71), poważa to, że z jednego rozwiązania można uzyskać drugie [1]? Bez względu na wyjaśnienie tego dylematu, wydaje się, że w rozdz. 3 zawarto wartościowe elementy przede wszystkim w pkt. 3.5. Związane są one ze wykazaniem różnic między dwoma podejściami. Niemniej jednak, nieco brakuje tutaj głębszej analizy różnic w złożoności obliczeniowej, a także w odniesieniu do literatury.
3. Rozdział 4. Deklarowany wkład Doktoranta dotyczy pkt. 4.4 i zdefiniowania odległości edycyjnej grafów, dla skierowanych grafów acyklicznych (DAG). Jest to krótki podrozdział i zawiera w sumie dość autorską interpretację odległości edycyjnej dla wybranego rodzaju grafów (DAG), wraz z odpowiednią adaptacją jednego algorytmu służącego do jej obliczenia (Algorytm 3, s. 104). Pewien niedosyt budzi tutaj brak głębszej analizy zarówno algorytmów, jak i porównania do innych, alternatywnych miar.
4. Rozdział 5. Autor twierdzi, że jego dokonanie to pełna zawartość tego rozdziału oraz że: „*opracowano metody mierzenia podobieństwa procesów ETL z wykorzystaniem odległości edycyjnej na drzewach i grafach*" (s. 7). Nie do końca wiadomo, co oryginalnego jest w owej metodzie opisanej w rozdz. 5? Zarówno miary, jak i algorytmy wydają się nie być w dużej mierze autorstwa Doktoranta. Autor jednak wymienia

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Katedra Sztucznej Inteligencji
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
T: +48 71 320 24 54
www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl
REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434

trzy propozycje własnych algorytmów (s. 127), ale opisuje je jedynie pojedynczym zdaniem (brak bliższego lub precyzyjnego opisu), np. *Hungarian-A*-top* przetwarza komponenty *w kolejności topologicznej*, przy czym nie do końca wiadomo, co to jest „*kolejność topologiczna*"? *Hungarian-A*-cut* ogranicza *przestrzeń poszukiwania do komponentów znajdujących się na zbliżonym położeniu w grafie*. Tutaj także nie wiadomo, co to jest „zbliżone położenie"? Niemniej jednak, bez wątpienia dokonaniem Autora jest eksperymentalna ocena algorytmów dokonana na trzech rzeczywistych zbiorach danych[1].

5. Rozdział 7. Według Autora, jego wkładem jest pełna treść tego rozdziału, a w szczególności metoda wnioskowania o brakujących powiązaniach pochodzenia danych przy użyciu dwuetapowego modelu bazującego na architekturze transformera (s. 7). W istocie, proces wnioskowania został sprowadzony do wnioskowania parami o tym, czy poszczególne elementy struktury danych wykorzystywanych w procesach ETL są ze sobą związane (jedna pochodzi od drugiej) na podstawie analizy tekstowych metadanych z nimi związanych. Oznacza to, że pochodzenie to pojedynczy krok transformujący dane. Pomimo różnych zastrzeżeń metodologicznych do tej części (patrz pkt. V), jest to niewątpliwie wartościowe dokonanie Autora wyrażone poprzez zbudowanie odpowiedniego klasyfikatora binarnego oraz propozycję przygotowania danych do uczenia (identyfikacja par, selekcja i filtrowanie kandydatów). Dodatkowo, walidacja zaproponowanego klasyfikatora została przeprowadzona na dużym zbiorze danych rzeczywistych pochodzących z firmy *Informatica*, co uwiarygadnia zaprezentowane wyniki.

Powyższa lista osiągnięć (oraz cała treść monografii) jest momentami nieco luźno ze sobą związana, aczkolwiek wszystkie dotyczą procesów ETL traktowanych jako wiele powiązań (przejść), które odpowiadają połączeniom (krawędziom) w grafach i drzewach oraz związkom semantycznym pomiędzy strukturami danych. Jest to absolutnie wystarczający poziom spójności badanych zagadnień.

Pomimo pewnych niedociągnięć (patrz także pkt. V) recenzowana monografia stanowi łącznie nowe, ważne i wartościowe osiągnięcie naukowe, w którym generalnie zastosowano odpowiednie, formalne i eksperymentalne metody badawcze.

IV. Treść rozprawy, elementy ogólnej oceny

Rozprawa została napisana w języku polskim, jest bardzo obszerna (232 strony), ale ogólnie zrozumiała. Treść składa się z 8 rozdziałów oraz bibliografii zawierającej 233 pozycje, w tym trzy Doktoranta.

Generalnie Autor zajął się kilkoma problemami, dla których dość szeroko opisał różne elementy najprawdopodobniej zaczerpnięte z literatury (brak jest

[1] Opis zbiorów mógłby być nieco lepszy (s. 127).

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Katedra Sztucznej Inteligencji
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
T: +48 71 320 24 54
www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl
REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434

wystarczającej liczby odpowiednich odesłań), zaś opis swoich własnych badań zbyt zredukował. Ów brak jest najbardziej widoczny w przypadku opisu algorytmów w rodz. 5.

Jedną z głównych kontrowersji w pracy jest niemal brak porównania z innymi rozwiązaniami znanymi z literatury. Oznacza to, że Autor bada własności swoich metod / modeli, ale nie wiadomo, czy są one lepsze lub gorsze od innych. Najbardziej brakuje tego w rodz. 7 (porównanie jest śladowe), zaś w rozdz. 3 jest to wykonane, ale na niewystarczającym poziomie.

Sensowność problemu rozważanego w rozdz. 7 jest trudna do oceny. Autor dysponował rzeczywistymi danymi, co jest bardzo dobrą okazją do pokazania, jak poważny jest rozważany problem. Wydaje się, że posiadane przez niego dane miały znane powiązania semantyczne między kolumnami, więc po co o nich wnioskować, skoro są znane? Jeżeli jednak nie wszystko było znane (wtedy co nie było? w jakich przypadkach?), to jaka była skala tej niewiedzy?

Ogólnie, w pracy brak jest konkretnych przykładów, zwłaszcza transformujących rzeczywisty problem w problem formalny lub problem uczenia maszynowego. W efekcie nie zawsze wiadomo, jaki rzeczywisty odpowiednik mają zdefiniowane lub opisane pojęcia i relacje. Nie ułatwia to zrozumienia i oceny całości dokonań.

Monografię można ocenić jako 5 pojedynczych dokonań, które nie do końca zostały dopracowane, zwłaszcza pod kątem właściwego opisu i analizy nowego wkładu. Owa mnogość pojedynczych dokonań daje jednak podstawę do stwierdzenia, że łącznie osiągnięcie Autora jest wartościowe i zasługuje na pozytywną ocenę.

Chciałbym także podkreślić, że poniższe, liczne pytania i uwagi nie są świadectwem niskiej jakości rozprawy, ale w dużej mierze wynikają z mojego zainspirowania się potencjalnie praktycznym odniesieniem zaprezentowanych badań. W szczególności pokazanie, że porównywanie struktur drzewiastych i grafowych oraz wnioskowanie o nich jest ważne w praktycznym utrzymywaniu i zarządzaniu wielkimi systemami, jest bardzo cennym ukierunkowaniem Autora i dowodem na sensowność koncepcji doktoratów wdrożeniowych.

V. Problemy, pytania i uwagi dyskusyjne

Problemy i pytania

1. Ze względu na braki odesłań do literatury trudno oddzielić to, co jest dokonaniem Doktoranta, od tego, co jest wzięte z literatury. Przykładowo, we wstępie do odległości edycyjnej grafów (pkt. 4.2.1) mamy odesłanie do „przeglądu systematyzującego" [119], ale nie wiemy, co z tego, co dalej Doktorant przedstawia, jest w istocie tym samym, co w owym przeglądzie – doktoracie z Włoch? Jedyną pomocą może być deklaracja „oryginalnego wkładu" Doktoranta zawarta w pkt.

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Katedra Sztucznej Inteligencji
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
T: +48 71 320 24 54
www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl
REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434

1.5. Wtedy, treść pkt. 4.2.1 pozostaje poza owym „wkładem". Pomijanie odesłań nie wydaje się być jednak dobrą praktyką.

2. Brak jest precyzyjnej definicji pochodzenia danych, która powinna być w pkt. 2.3.1, zwłaszcza w odniesieniu do zastosowania w ramach procesów ETL[2]. W efekcie nie bardzo wiadomo czym jest węzeł drzewa lub grafu? To „kolumna", czyli cecha, czy może pojedyncza wartość (instancja) tej kolumny? Autor swobodnie używa obu pojęć, co jeszcze bardziej komplikuje zrozumienie. W procesach ETL różne wartości końcowe (instancje) mogą mieć różne pochodzenie, np. suma zakupów klienta #3, w 01.2026 pochodzi ze sprzedaży online (jedne struktury) zaś w 02.2026 z tradycyjnego sklepu (inne struktury), co oznacza także inne pochodzenie.
3. Doktorant z mgliście zdefiniowanego pochodzenia przechodzi płynnie do drzew (czym jest drzewo w ETL?), co w efekcie prowadzi go do problemu porównywania drzew. Jako przykład podaje on struktury XML czy JASON. Rzeczywiście są to struktury drzewiaste, ale ich praktyczne zastosowania zawierają także elementy bardziej złożone, m.in. poprzez odesłania czy powielenia co w istocie zaburza drzewiasty charakter owych struktur. Warto zauważyć, że struktura drzewiasta często nie najlepiej odpowiada rzeczywistości. Zamodelowanie naturalnej relacji typu *m:n* poprzez drzewo jest zawsze kompromisem. Autor w istocie nie podaje żadnego przykładu, ilustrującego sens i potrzebę rozwiązania problemu podobieństwa drzew, w tym także tego, czy miary odległości edycyjnej dalej opisywane mają sensowne stosowanie. Nie jest to zasadniczy zarzut, ale skoro sam Autor odwołuje się do zastosowań (co nie było niezbędne), więc dobrze by było to lepiej powiązać. W rozdziale 3 nie dowiemy się także, czym są etykiety węzłów, czy jaka intuicja stoi za zmianą etykiety (def. 3.14) w odniesieniu do zastosowania w ETL.
4. Autor przedstawił Algorytm 1 (s. 79), przy czym nie porównał go do żadnego innego algorytmu znanego z literatury, w tym np. do wspomnianego przez siebie algorytmu RTED, którego nieco kompletniejsza analiza jest zawarta w rozszerzeniu pracy cytowanej jako [48] (arxiv), czyli w *ACM Transactions on Database Systems*, 40(1), 2015. Ponieważ ten algorytm zawarty jest w pkt. 3.6, który nie został wymieniony jako osiągnięcie Autora (patrz pkt. 1.3.1), więc chyba należy traktować go jako ilustrację rozwiązań z literatury – której publikacji? W takim przypadku zasadne jest pytanie o spójność z niewskazanym opisem literaturowym. Taka sama uwaga dotyczy algorytmu 2, umieszczonego w pkt. 4.3.3, który nie był wymieniony jako „wkład Autora".

[2] Mamy wprawdzie Def. 7.1, ale dotyczy ona tylko rozdz. 7 i nie za bardzo cokolwiek wyjaśnia w poprzedzających rozdziałach.

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Katedra Sztucznej Inteligencji
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
T: +48 71 320 24 54
www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl
REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434

5. W takim razie zasadnym pozostaje także pytanie o autorskość Algorytmu 3, który został umieszczony w pkt. 4.4.4, czyli w ramach zadeklarowanego „wkładu" Doktoranta. Dodatkowo, opisy algorytmów 1-3 mają drobne uchybienie: warunek pętli „dopóki nie znaleziono rozwiązania" jest mało precyzyjny.
6. Brak eksperymentalnej weryfikacji algorytmów, w tym zwłaszcza porównania do innych.
7. Brak bardziej precyzyjnej lub formalnej definicji pojęcia *przepływu*. Praca zawiera jedynie bardzo ogólne i nieprecyzyjne opisy i przykłady. W efekcie mamy niezdefiniowane pojęcia „kolumny" (s. 23-24). Autor wprawdzie dużo dalej, tj. w rozdz. 5 podał definicję „modelu MetaDexa" (5.1, s. 109-110), ale nie wiadomo w nim czy:
   a) ów *model* jest tym samym co *przepływ*?
   b) Co to jest *kolumna*?
   c) Co to jest *komponent* – wierzchołek grafu komponentów? Nie podano przykładu. Z Rys. 5.1 i związanego z nim opisu trudno to jednoznacznie zrozumieć. Czy to jest rodzaj operacji (np. *złączenie*), konkretna operacja, czy może cecha (jedna, wszystkie?) przetwarzana w ramach operacji?
   d) Co to jest *pole* – wierzchołek *grafu pól*? Prócz tego, że pole należy do komponentu, z definicji nie dowiemy się czym jest pole. Wprawdzie przed definicją pojawia się informacja, że „*model MetaDexa posiada dodatkową strukturę pól oraz pól kontrolnych*" (s. 109), ale to niespecjalnie wyjaśnia to, czym jest pole?
   e) Co to jest krawędź między polami i czym się różni od krawędzi między komponentami?
   f) Czy pole należy zawsze do jednego komponentu, czy może należeć do wielu lub żadnego?
   g) W definicji (chyba to jej część?) pojawia się także trudno zrozumiałe pojęcie *drzewo wyrażeń* zawarte w etykietach pól (s. 110) – co to jest? Etykiety pól to drzewa?
   h) Gdyby dla przykładowego modelu MetaDexa na poziomie kolumn (czy są inne poziomy?) z Rys. 5.1 podano odpowiadające mu grafy komponentów i pól, to może cała definicja byłaby jaśniejsza.
   i) W definicji modelu nie pojawia się pojęcie *operacji*, które z kolei jest używane dalej i nawet posiada typy (np. s. 112), co dodatkowo komplikuje zrozumienie całości modelu. Czy typy operacji mają jakiś związek z etykietami grafów?
8. Badania eksperymentalne w rozdz. 5 mają miejscami niejasny opis. Generalnie nie bardzo wiadomo, co to jest *koszt*, gdyż wcześniej niezbyt precyzyjnie podano, że operacje mają ważony koszt, ale nie do końca wiadomo jaki? Czy dodanie jednego węzła to koszt=1? Czy zawsze? Na s. 128 mamy tylko, że „*koszt usunięcia, dodania lub zamiany etykiety*

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Katedra Sztucznej Inteligencji
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
T: +48 71 320 24 54
www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl
REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434

krawędzi ustawiono na 10". Dodatkowo, w tabelach chyba podano średni koszt (w Tab 5.1 i 5.2). Jednak nie wiadomo w ramach jakiego zbioru owe średnie były obliczane, tzn. ile było klastrów i jakich oraz ile w efekcie porównań w parach zostało obliczone? Dodatkowo, jeżeli koszt operacji na jednej etykiecie to 10, a średni koszt w Tab. 5.1 dla większości algorytmów nie przekracza 5, czy oznacza to, że etykiety bardzo rzadko w ogóle były zmieniane przy edycji grafów i niemal nie było żadnych innych operacji edycyjnych? Nie wydaje się to wiarygodne, zwłaszcza, że koszty w Tab. 5.2 mają wartość ok. 300. Nie wiadomo także, co zawiera Tab. 5.2. Dla jakich wybranych grafów losowych podano w niej wyniki? Czy chodzi o losowanie grafów (ilu?) i pomiar ich podobieństwa do wszystkich innych? W efekcie dla ilu porównań par podano dane? W pkt. 5.1.4 mamy formalną definicję problemu wprowadzającą próg τ. Jaka wartość tego progu przyjęto w eksperymentach i jakie miało to konsekwencje?

9. Dlaczego w (Tab. 5.1, 5.2) nie podano wartości Δ dla niektórych przybliżonych algorytmów (H-A*-top, H-A*-cut, DAG-Edit)?
10. Rozdział 6 to dłuższe wprowadzenie, także historyczne, do przetwarzania języka naturalnego, które jest potrzebne w ograniczonym stopniu i niezbyt wiele wnosi do całej, bardzo obszernej monografii. Prawdopodobnie mogłoby ono być skrócone do jednej strony w ramach wprowadzenia w rozdz. 7. Warto przy tym pamiętać, że rozdział ten nie został wskazany jako „wkład" przez Autora.
11. W rozdz. 7 Autor opisuje problem odtwarzania pochodzenia danych, ograniczając się w istocie tylko do dopasowania schematów (struktur danych) i twierdzi „*Tradycyjne metody zakładają porównanie wszystkich możliwych par elementów*" (s. 158). Dlaczego porównywanie par (i ewentualne uprzednie blokowanie lub filtrowanie kandydatów) jest jedynym i/lub najlepszym podejściem. Dlaczego nie można tego potraktować jako problem wieloetykietowy (może czasami wieloklasowy?) identyfikacji dla danego elementu innych (innego) elementu, od którego pochodzi? Co więcej, uczenie takiego problemu oznacza zwykle ogromne niezbalansowanie danych uczących (par niepasujących jest zdecydowanie dużo więcej niż pasujących)[3].
12. Z rozdz. 7, podobnie jak w pozostałych, zabrakło pojedynczego przykładu uczącego, co pomogłoby zrozumieć rozważany problem, który nie został wystarczająco precyzyjnie zdefiniowany. W efekcie nie bardzo wiadomo, jaka jest relacja między danymi źródłowymi a docelowymi, 1:1, 1:n, m:n? Wydaje się, że w praktyce jest to relacja m:n, jednak treść rozdz. 7 raczej sugeruje 1:1 (model jest binarny), co może być poważnym ograniczeniem stosowalności zaproponowanych rozwiązań. Dodatkowo, czy relacja rozważana między kolumnami jest symetryczna, tzn. kolumna A pochodzi od B i równocześnie B od A?

[3] Autor sam zauważył owo niezbalansowanie w swoich danych, tj 1:90.000 (s. 173).

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Katedra Sztucznej Inteligencji
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
T: +48 71 320 24 54
www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl
REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434

Rozważania z poprzednich rozdziałów sugerują, że nie jest ona symetryczna.

13. Czy w istocie interpretację problemu z rozdz. 7 można traktować jako podejście *binary relevance* dla problemu wieloetykietowego? Jeżeli tak, to czy nie byłoby wartościowe pokazanie danych statystycznych (zagregowanych rozkładów, zbalansowania klas, …) oraz odpowiednich dla tego problemu miar, np. micro / macro F1?
14. Nie w pełni jest jasna konfiguracja eksperymentów w rozdz. 7. Nie bardzo wiadomo, jakie transformacje między schematami były uwzględniane. Czy „*5673 źródeł*" oznacza tyleż firm? Jeżeli tak, to mamy średnio prawie 14 schematów na firmę. Czy z każdego schematu danej firmy były robione transformacje na każdy inny schemat tej firmy? Czy mogły być kolumny w danym schemacie, które nie miały żadnego pochodzenia w innym schemacie w rozważanej parze? Jaka była liczność zbioru uczącego i treningowego w różnych scenariuszach A, B, C, między- i wewnątrz projektowe? Jaki model klasyfikatora był w istocie użyty? Czy dokonano optymalizacji hiperparametrów – jakie były hiperparametry? Na s. 176 Autor wspomina o czasie trwania pojedynczych sesji, ale nie wiadomo, co to jest 'sesja'?

Uwagi do dyskusji

15. Ważną cechą rzeczywistych systemów jest ich dynamika, czyli zmienność struktur oraz całych procesów ETL. W związku z tym, czy Autor rozważał np. problem wersjonowania?
16. Z pojęciem *data lineage* wiąże się wiele innych problemów i zagadnień, np. detekcja i korekcja błędów, optymalizowanie / poprawa przepływów czy transformacji, spójność danych, czy analiza wpływu zmian. Zostało to częściowo wymienione w pkt. 2.3.4 (s. 25-26).
17. Autor twierdzi, że „*generowanie trudnych negatywów (ang. hard negatives) stanowi kluczowy komponent skutecznego uczenia modeli*" (s. 159). Jest to nieco kontrowersyjne stwierdzenie. Negatywne, „*semantycznie zbliżone*" nie muszą być dobre dla dobrego wytrenowania klasyfikatora. Istnieje sporo badań i rozwiązań, które pokazują, że przykłady negatywne nie powinny być zbyt podobne do pozytywnych, gdyż wtedy model może w ogóle nie nauczyć się danego problemu lub łatwo się przeuczyć i nie być w stanie właściwie generalizować problem. Innym przykładem odnoszącym się do tej kwestii są subiektywne zadania w NLP, np. detekcja emocji, dla których ten sam tekst jest odbierany przez różne osoby w różny sposób. Wtedy dla danego tekstu mamy sprzeczne etykiety wyjściowe. A przecież przykład negatywny (tekst z inną emocyjną etykietą) jest w takim przypadku najbardziej podobny semantycznie, bo jest taki sam. Z tego powodu negatywne przykłady zwykle ustala się jako *odpowiednio bliskie / dalekie*. Dodatkowo Autor nie przebadał, czy owe

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Katedra Sztucznej Inteligencji
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
T: +48 71 320 24 54
www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl
REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434

„generowanie przypadków negatywnych" jest dobrym/najlepszym rozwiązaniem.

18. Czy w problemie odtwarzania pochodzenia, Autor zastanawiał się nad kwestią alternatywnych pochodzeń (różnych możliwych): dany element może być utworzony z różnych innych elementów i każde takie pochodzenie jest właściwe. Przykładowo, płeć można wywnioskować zarówno z PESEL-u, jak i z imienia osoby.
19. Autor w rozdz. 7 skupia się na tekstowych opisach struktur danych (metadanych). Czy rozważane było wykorzystanie także zawartości kolumn (instancji) a nie tylko ich metadanych?
20. Zamiast prowadzić eksperymenty na dużym zbiorze danych, być może lepiej było ograniczyć się do jakiegoś podzbioru, ale wykonać szersze badania, np. różne modele i hiperparametry, walidację krzyżową, itd.

Uwagi formalne i redakcyjne

21. Definicja drzewa (3.5) nie zawiera założenia o skończoności zbioru wierzchołków (zawarto to w zdaniu poprzedzającym), aczkolwiek w kontekście zastosowań nie jest to specjalnie istotne. Nie precyzuje ona także, ile może być korzeni, co może mieć konsekwencje dla dalszych rozważań, zwłaszcza w odniesieniu do definicji lasu (3.9), która odnosi się do *drzew ukorzenionych*. Jednak wcześniej nie sprecyzowano, co to jest *drzewo ukorzenione*? Mające co najmniej jeden korzeń? Dokładnie jeden korzeń?
22. Nieco dziwne sformułowania, np. „*A wczesniej pisałeś Na rysunku 2.2 przedstawiono przykładowy proces ETL w CDI*", s. 13., „*w tekście w ogóle nie odnosisz się do tego algorymu; musi być powołanie w tekście na ten algorytm, najlepiej z jakims komentarzem*" (s.98)
23. Autor powiela Rys. 2.7 w rozdz. 7 jako Rys. 7.1, po czym odnosi się do wcześniejszego rysunku: „Rysunek 2.7 przedstawia..." (s. 148). Po co w takim razie jest powielenie rysunków?

VI. Zewnętrzna ocena pracy - publikacje

Wyniki rozprawy zostały opublikowane w 3 artykułach czasopismowych, tj. w: (1) *Schedae Informaticae* 2023, 20 pkt. (nieindeksowane w WoS czasopismo UJ), (2) *Proceedings of the VLDB Endowment*, 2025, IF=2,6, 20 pkt.[4], oraz (3) *Annals of Combinatorics*, IF=0,7, 100 pkt (ukazał się po złożeniu rozprawy już w 2026 roku). Nie jest to imponujący, ale akceptowalny zbiór publikacji. Oznacza on również, że

[4] Jest to w istocie publikacja pokonferencyjna warsztatów "*AIDB 2025: The 6th Applied AI for Database Systems and Applications Workshop*" przy konferencji VLDB 2025 (szkoda, że Doktorant o tym nie napisał, co w efekcie wymagało odpowiedniej kwerendy z mojej strony). Sama konferencja VLDB ma ranking CORE A* i 200 pkt. oraz jest jedną z najlepszych i najstarszych konferencji obejmujących tematykę hurtowni i przetwarzania danych.

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Katedra Sztucznej Inteligencji
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
T: +48 71 320 24 54
www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl
REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434

zewnętrzna ewaluacja osiągnięć Doktoranta została dokonana w ograniczonym, ale wystarczającym zakresie.

Ogólnie, w serwisie Scholar zarejestrowano 11 prac Doktoranta i były one cytowane 23 razy (wraz z autocytowaniami), h-indeks=3, zaś w serwisie Scopus znajduje się 5 prac, z których tylko dwie (z 2018 i 2019 roku) były cytowane 7 razy (bez autocytowań), h=2. Wszystkie prace były publikowane albo na konferencjach albo w czasopismach (także preprinty) o umiarkowanej renomie.

VI. Podsumowanie i ocena końcowa rozprawy

Podsumowując, należy stwierdzić, że rozprawa jest łącznie ciekawa, bardzo obszerna, dojrzała i dotyczy użytecznej tematyki zarządzania procesami przetwarzania dużych i złożonych zbiorów danych. Jej charakter można określić jako teoretyczno-eksperymentalny. Bez wątpienia stanowi ona istotny wkład w dyscyplinę „informatyka techniczna i telekomunikacja" w „dziedzinie nauk inżynieryjno-technicznych". Osiągnięcia wymienione powyżej w pkt. III są łącznie nowe i wartościowe. Ich ocena zawarta w pkt. IV, pomimo uwag i pytań z pkt. V jest wg mnie jednoznacznie pozytywna.

Całość świadczy pozytywnie o dobranym warsztacie badawczym i sporych i różnorodnych umiejętnościach Doktoranta. Treść jest metodologicznie akceptowalna, pomimo pewnych zastrzeżeń. Praca, pomimo istotnego charakteru teoretycznego, ma bardzo duże znaczenie praktyczne, gdyż zarządzanie wielkimi, dynamicznymi zasobami jest i pozostanie dużym wyzwaniem.

W związku z powyższym stwierdzam, że opiniowana rozprawa doktorska mgra Macieja Brzeskiego spełnia wymagania stawiane w obowiązujących przepisach ustawy o stopniu naukowym doktora i wnoszę o dopuszczenie jego Autora do publicznej obrony.

Przemysław Kazienko

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Katedra Sztucznej Inteligencji
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
T: +48 71 320 24 54
www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl
REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434