Gliwice, 10 stycznia 2026

prof. dr hab. inż. Dariusz Mrozek
Katedra Informatyki Stosowanej
Politechnika Śląska w Gliwicach
ul. Akademicka 16
44-100 Gliwice

RECENZJA

rozprawy doktorskiej dla Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja działającej w Politechnice Wrocławskiej

Tytuł rozprawy: Fixing data quality issues in CMDB in IT and OT by using machine learning algorithms. Forecasting for IT procurement

Autor rozprawy: Szymon Niewiadomski

1. Zagadnienie naukowe rozpatrzone w pracy

Przedstawiona przez Pana Szymona Niewiadomskiego rozprawa doktorska jest poświęcona poprawie jakości danych w bazach danych zarządzania konfiguracją (CMDB) i predykcyjnym strategiom nabywczym sprzętu komputerowego w przedsiębiorstwie poprzez zastosowanie odpowiednio zaprojektowanych metod wstępnego przetwarzania danych, modeli sztucznej inteligencji (AI) i algorytmów ewolucyjnych. Główne cele rozprawy koncentrują się wokół zagadnień poprawy procesów walidacji rekordów wprowadzanych do bazy CMDB, okresowej weryfikacji jakości danych w tych bazach, wykrywania błędów strukturalnych w grafie bazy CMDB, a także optymalizacji procesu nabywania zasobów infrastruktury IT. Zarówno cele pracy, jak i motywacja prowadzonych badań w tym obszarze zostały sformułowane w sposób jasny. Rozprawy ma charakter wdrożeniowy, Autor:

- zaproponował swoje usprawnienia dla reprezentacji danych, ich przetwarzania i poprawy ich jakości,
- dla potwierdzenia słuszności przyjętych rozwiązań przeprowadził badania eksperymentalne, które pozwoliły zweryfikować, iż opracowane algorytmy i rozwiązania mogą być pomocne do rozwiązania ww. procesów.

Po lekturze materiału wydaje się, iż założone cele rozprawy udało się Autorowi osiągnąć.

2. Umiejscowienie problemu rozpatrywanego w rozprawie w kontekście światowej literatury

Analiza światowej literatury i bieżącego stanu wiedzy w omawianym obszarze przedstawione przez Autora w rozdziale 1.10 wskazują, że problematyka jakości danych w bazach CMDB jest tematem istotnym. Niestety analiza ta jest bardzo zdawkowa, a cytowane prace mało świeże – większość przytoczonych prac ma powyżej 10 lat, a brakuje prac z ostatnich 5 lat (zwracam uwagę, iż mówię tu o rozdziale 1.10). Brakuje porządnego przeglądu literatury przedmiotu, np. z podziałem na kategorie problemów jakościowych. Powoduje to, że trudno spozycjonować wyniki osiągnięte przez Autora rozprawy na tle światowej literatury. Tym bardziej, że osiągnięte wyniki nie zostały bezpośrednio skonfrontowane z wynikami prac pokrewnych. A przecież problemy jakości danych i ich czyszczenia są obecne w literaturze od 30 lat i nie dotyczą samych tylko baz CMDB, ale powstała cała gama algorytmów poprawy jakości danych przy okazji realizacji np. procesów ETL do hurtowni danych. I być może należałoby sięgnąć do tej grupy prac, aby dobrze umiejscowić swoje rozwiązanie na tle tych rozwiązań. Autor wskazuje co prawda na wady istniejących metod, które są intuicyjnie prawdziwe, ale brakuje dobrego przeglądu prac i podparcia eksperymentalnego. W spisie literatury znajduje się 69 pozycji literaturowych, a i tak nie wszystkie z nich są cytowane w tekście.

3. Poprawność rozwiązania i przyjętych założeń

Na początku realizacji rozprawy Pan Szymon Niewiadomski zdefiniował kilka zadań, do których realizacji dążył w swoich pracach badawczych. Dotyczyły one możliwości realizacji wydajnego zarządzania jakością danych w bazach CMDB, obejmującej kompletność, dokładność i zgodność danych, wskazując na pojawiające się błędy w postaci zduplikowanych rekordów, brak unikalnych identyfikatorów, błędy typograficzne, niewłaściwe użycie pól wprowadzania danych, błędy wynikające z niedokończonych wdrożeń, czy niewłaściwy status cyklu życia systemu. W swoich pracach Autor sięgnął do rozwiązań bazujących na automatycznym generowaniu cech wykorzystując tokenizację n-gramową, dyskryminację opartą na gęstości, filtracji cech, a także klasyfikatorach bazujących na metodach sztucznej inteligencji, obejmujących m.in. sztuczne sieci neuronowe. W stosunku do optymalizacji strategii nabywania zasobów infrastruktury IT oparł swoje prace na algorytmach genetycznych. Na podstawie lektury rozprawy można domniemywać, iż postawione w rozprawie zagadnienia zostały rozwiązane w sposób właściwy. Autor osiągnął to poprzez: 1) identyfikację słabości istniejących metod poprawy jakości opartych na zbiorach reguł, 2) opracowanie własnych usprawnień i algorytmów lub użycie istniejących algorytmów, 3) badania eksperymentalne weryfikujące przydatność opracowanych metod z użyciem różnych zbiorów danych zawierających poprawne i niepoprawne dane. Wg. przedstawionych w pracy wniosków, wyniki przeprowadzonych przez Autora rozprawy badań potwierdziły, iż założenia przyjęte podczas opracowania autorskich metod były słuszne i uzasadnione. Jak wspomniałem wcześniej, brakuje jednak porównania osiągniętych wyników z wynikami istniejących i popularnych narzędzi powszechnie używanych przez specjalistów prowadzących podobne analizy danych i procesy klasyfikacyjne.

4. Oryginalność rozprawy i samodzielny dorobek Autora

Przedstawiona rozprawa mogłaby stanowić dobre uzupełnienie bieżącego stanu wiedzy światowej w zakresie prowadzenia wydajnych poprawy jakości baz danych konfiguracyjnych, gdyby w rozprawie znalazły się jakiekolwiek odniesienia do światowych wyników badań w tym obszarze lub w obszarach pokrewnych, np. bardziej uogólnionych (np. poprawa jakości danych w bazach danych). Wydaje się, że oryginalność polega tu na opracowaniu własnych usprawnień w obszarze algorytmów selekcji cech (dyskryminacja i informatywność), algorytmu wykrywania błędów strukturalnych w bazie CMDB wykorzystując komplementarny naiwny klasyfikator Bayesa (Complement Naive Bayes, CNB), umiejętnym wykorzystaniu innych zdobyczy nauki i odpowiedniej ich integracji w kontekście baz zarządzania konfiguracją. Wkład własny Autora w stosunku do istniejących metod został jednak dość słabo uwypuklony. Trudno też stwierdzić, iż w rozprawie Autor przeprowadził proces wnikliwej (eksperymentalnej) oceny swoich rozwiązań, co jest istotną wadą rozprawy.

5. Zawartość rozprawy i jej struktura

Realizując pracę Pan Szymon Niewiadomski wykazał niezłe opanowanie umiejętności przedstawiania opracowanych przez siebie rozwiązań i uzyskanych wyników badań eksperymentalnych. Same idee zostały zaprezentowane w sposób dość sformalizowany, choć ten formalizm nie ułatwia wcale lepszego zrozumienia przedstawionych idei i w wielu miejscach mógłby być znacząco ograniczony na rzecz lepiej przeprowadzonych badań eksperymentalnych.

Praca zawiera 4 główne rozdziały oraz załączniki:

[R1] Rozdział 1 rozprawy wprowadza czytelnika w problematykę baz CMDB, istotę zarządzania konfiguracją, pojemnością i żądaniami, a także grafową reprezentację zawartości baz CMDB, wraz ze stojącą za nią motywacją i formalizmem matematycznym. Wskazuje cele przemysłowe realizowanej rozprawy w ramach doktoratu wdrożeniowego, a także obszary operacyjne, w ramach których przygotowano rozwiązania algorytmiczne. Przedstawia ogólne diagramy przepływu sterowania określając jednocześnie miejsca zastosowania metod sztucznej inteligencji w procesach walidacji danych wejściowych do systemów CMDB, miesięcznej kontroli jakości istniejących w CMDB danych, obsługi błędów wyłapanych przez modele AI. Autor przeprowadził w nim także pobieżny przegląd literatury wskazując niedoskonałości przytoczonych metod bazujących na regułach.

[R2] Rozdział 2 jest poświęcony problematyce czyszczenia danych. Przedstawiono w nim m.in. opracowane przez Autora, oparte na uczeniu maszynowym, podejście do comiesięcznego sprawdzania poprawności rekordów bazy CMDB (automatyczna ekstrakcja cech oparta na tokenizacji n-gramów, filtracja cech bazując na niskiej częstości występowania i zmienności wartości, możliwościach dyskryminacyjnych, informatywności). W rozdziale przedstawiono także algorytmy wykrywania podejrzanych rekordów bazy CMDB oparte na analizie odległości między rekordami – globalne sprawdzanie wszystkich par rekordów, lokalne oparte na metodzie najbliższych sąsiadów i regresji jądrowej. Wreszcie przedstawiono bazujący na metodzie CNB algorytm wykrywania błędów strukturalnych w grafie bazy CMDB obejmujących brakujące węzły grafu (elementy CI) i krawędzie, naruszenia konwencji nazewnictwa, a także niepoprawne powiązania w grafie. Wskazano zalety

zaproponowanego podejścia, przy czym brakuje w nim porównania opracowanych metod do metod literaturowych, a wykorzystane w eksperymentach zbiory słabo korelują z danymi firmy, w ramach której realizowano doktorat wdrożeniowy.

[P3] Rozdział 3 poświęcił Autor problematyce optymalizacji strategii nabywania sprzętu komputerowego w dużych przedsiębiorstwach. Przy użyciu algorytmu genetycznego optymalizowano tu strategię pojedynczych, dużych zakupów oraz mniejszych i częstszych zakupów biorąc pod uwagę cenę, zapotrzebowanie, stopy procentowe i czas realizacji. W badaniach eksperymentalnych zweryfikowano wyniki procesów optymalizacyjnych dla różnej liczby pokoleń algorytmu genetycznego.

[P4] Rozdział 4 podsumowuje aspekty wdrożeniowe opracowanych rozwiązań w kontekście wymagań zarządzania zmianą, wewnętrznej implementacji, architektury zorientowanej na usługi, strategii konserwacji i wsparcia, a także wewnętrznych procedur spółki Gaz-System.

Struktura rozprawy jest dość dobra, natomiast nie udało się Autorowi zachować pełnej spójności treści rozprawy. Podam kilka przykładów, choć jest ich więcej. W rozdziałach 1.10.3 i 1.10.5 Autor pisze o „recursive trust indicators methodology" jako własnym podejściu w detekcji anomalii w danych z baz CMDB i, wydawałoby się, zapowiedź tego, co zostanie przedstawione w rozdziale 2. Niestety taka nazwa metodologii nie pojawia się już więcej w rozprawie. Podobnie jest ze spójnością matematyczną – np. grecka litera eta η jest użyta trzy razy w różnych kontekstach (takich przykładów jest więcej). Zamienne użycie określenia Rajski vs. Jaccard (nagłówek rozdz. 2.1.7 vs. treść rozdziału i pracy), należałoby się zdecydować na jedno określenie i jego konsekwentne użycie. Stosowanie synonimów i analogów, szczególnie dla własnych dokonań utrudnia zrozumienie treści czytelnikowi, a czasami nawet lepszą identyfikację dokonań Autora. Doceniam natomiast przedstawione w rozdziale 2 ogólne przedstawienie kroków algorytmów z ich bardziej szczegółowymi opisami w kolejnych podrozdziałach.

6. Poprawność redakcyjna rozprawy

Od strony redakcyjnej praca jest napisana w dość dobrym stylu. Pojawiają się w niej jednak odwołania do nieistniejących rozdziałów, sekcji czy rysunków, jak również osierocone odwołania do literatury. Uważam, że takie błędy powinny być wychwycone przez Autora w procesie powtórnego czytania pracy przed jej ostatecznym złożeniem. Problematyka pracy jest jasna. Same idee zostały zaprezentowane w sposób dość sformalizowany, lecz brakuje dobrych przykładów na poparcie idei, szczególnie z obszaru realizowanego doktoratu wdrożeniowego. Oceny skuteczności rozwiązań dokonano dość pobieżnie, czasem z użyciem publicznie dostępnych zbiorów danych, choć przy realizacji doktoratu wdrożeniowego prosiłoby się o więcej przykładów pochodzących bezpośrednio z firmy (od trywialnych po te znacznie bardziej skomplikowane), dla której przygotowywane było to rozwiązanie.

7. Słabe strony rozprawy i jej główne wady

Przedstawione rozwiązania są ciekawe i dotyczą istotnych problemów poprawy jakości danych i powiązań pomiędzy rekordami w bazach CMDB. Mam jednak kilka uwag i pytań, na które odpowiedź chętnie bym poznał:

U1. Przedstawiony w rozdziale 1 przegląda metod jest bardzo szczątkowy i wybiórczy. Problem czyszczenia danych jest problemem znanym od lat. Jak przedstawione metody korespondują np. z metodami czyszczenia i walidacji danych znanymi z procesów ETL? Powstaje przy tym pytanie czy proces kontroli jakości i czyszczenia danych w systemach CMDB różni się znacząco od analogicznych procesów prowadzonych na rzecz innych domenowych baz danych czy hurtowni danych? Wiele metod dla tego obszaru powstało na przestrzeni ostatnich kilku dekad. Autor rozprawy kompletnie pomija te obszary informatyki. Dlaczego? i dlaczego nie dokonano takiego porównania?

U2. Brakuje przekonywujących dowodów, że przedstawione metody są lepsze niż metody bazujące na regułach (z rozdziału 1.10.2).

U3. Pomijając przegląda literatury z rozdz. 1.10, w całym rozdziale 1 brakuje odnośników do literatury, a wiele faktów jest przytaczana bez podparcia literaturowego – np. przypadek ilustracyjny ze str. 21 dot. organizacji z sektora ubezpieczeniowego. Skąd te dane? Ale problem jest szerzy, bo pojawia się tam wiele definicji i klasyfikacji, które bez podparcia literaturowego sugerują, że Autor rozprawy jest autorem tych pojęć i klasyfikacji lub, że wiedza ta została zagregowana przez generatywne modele AI.

U4. Niektóre skróty pojawiają się w pracy bez odpowiedniego rozwinięcia, np. pierwsze użycie ITSM, ITOM, ITAM.

U5. Proszę o interpretację wzoru 2.9, czym jest D jako granica całkowania, oraz jak ta definicja *delta^(j)* koresponduje z definicją podaną w Step 3 dwie strony wcześniej. Na jakiej podstawie Autor stwierdza, że następuje tak duża redukcja „M $\sim 10^4$ to M' $\sim 10^2$"?

U6. Nie jest dla mnie jasne, a z opisu to nie wynika, dlaczego we wzorach 2.10 i 2.11 w mianownikach przed sumą nie występują parametry h_C i h_E, jak we wzorach 2.6 i 2.7? Jeśli jest to celowe, to należałoby to uzasadnić.

U7. W rozdziale 2.1.4, pojawiają oznaczenia h_B i N_B, wcześniej Autor operował na h_E i h_C oraz N_E i N_C. Skąd się biorą te oznaczenia?

U8. Pod wzorem 2.15 pojawia się zapis $h_C(N_C)$ - jak go należy interpretować? czy h_C jest funkcją o argumencie N_C?

U9. Na początku rozdz. 2.1.5 brakuje mi motywacji dla prowadzenia tych rozważań w kontekście kroków algorytmu z rysunku 2.1. Czego dokładnie dotyczą te rozważania?

U10. Na str. 41, w sekcji *Rarity controls* pojawiają się pojęcia niezdefiniowane i nieużywane wcześniej, takie jak *min_df*, Pi_{min} – zapewne jakieś parametry użytych metod bibliotecznych. Czym jest *target-class share*? Następnie d_R na stronie 42.

U11. Brakujące odwołania, np. str. 44 cf. [?], str. 47 equation (??).

U12. We wzorze 2.33 skoro prawdopodobieństwo iloczynu jest iloczynem prawdopodobieństw, to zakładamy niezależność zdarzeń. Dlaczego możemy to założyć?

U13. Jak mają się wykresy funkcji f_A oraz f_B do gęstości f_C oraz f_E? proszę też podać interpretację zmiennych p oraz q we wzorze 2.34? Dlaczego wartości p i q nie zależą od indeksu j?

U14. Rozdz. 2.1.9, brakuje porównania wyników opracowanej metody z innymi metodami z grupy *feature selection* i wcześniejszego przeglądu literatury w tym zakresie.

U15. W rozdz. 2.1.10, jaki był cel przeprowadzonego eksperymentu? O tym należałoby napisać na początku tego rozdziału. W tym kontekście pojawia się również pytanie dotyczące zdania „Figure 2.6 illustrates the superiority ..." – brak stwierdzenia co porównujemy ze sobą.

U16. Nie rozumiem obliczeń przestawionych w sekcji *Economic Evaluation* na str. 50. Jak otrzymano wartość $1000 na wykrycie jednego błędu i skąd bierze się wartość $109? Zastosowano tu chyba jakieś skróty myślowe.

U17. Dlaczego w rozdziale 2.2.3 użyto zupełnie innego zbioru danych niż we wcześniejszych badaniach eksperymentalnych i jaki mają one związek z bazami konfiguracyjnymi CMDB?

U18. Złożoność obliczeniowa metod przedstawiona w Table 2.2 różni się od tej opisanej w tekście na górze strony 53. Dlaczego?

U19. Na stronie 54 przedstawiono wektor wag w^T – w tekście nie podano skąd się bierze liczba jego elementów, których cech dotyczą wagi i jak dobrano wartości wag.

U20. Dlaczego w rozdziale 2.2.3 przedstawiono wyniki tylko dla metody *All-Pairs Distance*, a nie przedstawiono ich dla pozostałych dwóch lokalnych metod? Nie porównano wyników działania wszystkich trzech podejść, ani jakości wyników, ani pod względem weryfikacji czasu ich wykonania. Zostajemy zatem z suchymi stwierdzeniami i teoretyczną analizą złożoności obliczeniowej. Wyników własnych nie porównano z wynikami prac pokrewnych (o tym już wspominałem).

U21. Na początku sekcji 2.3 Autor pisze, że sekcja ta i sekcja 2.4 są zorientowane głównie na implementację, podczas gdy mamy tu cały zestaw formalizmów opisujących algorytmy. Pisze Autor również o „reproducibility", podczas gdy w zasadzie nie ma tam kodów oprogramowania. Dlaczego Autor nie udostępnił kodów oprogramowania przez repozytorium GitHub, zważywszy, że mamy do czynienia z doktoratem wdrożeniowym?

U22. W tym rozdziale 2.3 znów pojawiają się te same zmienne o różnym znaczeniu, np. B jako budżet, B jako „random relabelings", w algorytmie 1 jako „optional permutations". Z pracy nie wynika, że jest to tożsame. Czy jest? Powstaje pytanie jak budżet B został ujęty w Algorytmie 1, gdzie i jak wprowadza ograniczenia?

U23. W rozdziale 2.3.3 w sekcji „Steps to transform the CMDB Graph" dostajemy znów ogólną wiedzę literaturową dotyczącą reprezentacji grafu, ekstrakcji cech i pozostałych kroków. Dobrze, ale znów brakuje odwołań do literatury. Po drugie, nie wiadomo jakiego ostatecznie wyboru dokonał Autor np. do reprezentacji grafu, selekcji cech, obsługi brakujących danych, normalizacji w swojej pracy i dlaczego poczynił takie wybory?

U24. W rozdziale tym nie sprecyzowano także, jaki był konkretny cel eksperymentu, którego wyniki przedstawiono w Tabelach 2.5 i 2.6. W jaki sposób użyty klasyfikator CNB był trenowany i testowany? Jakiego zbioru użyto, w jakich proporcjach?

U25. Na początku rozdziału 2.4 pojawia się stwierdzenie „The proposed algorithm ..." – nie wiadomo w zasadzie, o który algorytm chodzi, ponieważ rozdział 2.4 jest na tym samym poziomie pracy, co 2.1, 2.2 i 2.3, a każdy z nich przedstawia jakąś metodę lub algorytm. Poza tym Autor pisze, że jego model jest lekki, ale w rozprawie brakuje informacji o jego wielkości i ew. liczbie parametrów. Co do zaś wyjaśnialności, nie jest dla mnie jasne z lektury pracy w jaki sposób ją uzyskano i znów brakuje porównania z typowymi metodami stosowanymi w tzw. wyjaśnialnej AI (np. SHAP czy LIME).

U26. W rozdziale 3.2 na stronie 74, rysunek 3.1 nie pokazuje idei algorytmu, a jedynie jego dane wejściowe i wyjście.

U27. Opis algorytmu genetycznego przedstawionego w rozdziale 3.5 jest bardzo zdawkowy. Wygląda na użycie typowego algorytmu genetycznego, ale w opisie znów brakuje szczegółów. Brakuje dokładnego opisu funkcji przystosowania, nie wiadomo jakie są prawdopodobieństwa użycia operatorów krzyżowania i mutacji, nie wiadomo, ile było osobników w populacji i ile średnio tych populacji było. Nie wiadomo ile razy uruchamiany był algorytm (bo skoro to jest algorytm zrandomizowany, to powinien był być prawdopodobnie uruchamiany wielokrotnie, żeby osiągnąć jakieś uśrednione wyniki).

U28. Autor nie zinterpretował w tekście rozdziału 3.6 wyników eksperymentów przedstawionych na rysunkach 3.7, 3.8 i 3.9, a przedstawione wykresy są mało czytelne.

U29. Z rozdziału 3 nie dowiadujemy się, czy algorytm został rzeczywiście wdrożony. Poza tym nie wynika z niego jak Autor poradził sobie z problemem niepewności w przewidywaniu czasu dostawy, ani nie przedstawiono algorytmu dla takiej predykcji.

U30. W rozdziale 4.3 Autor pisze o architekturze systemu, ale nie została ona nigdzie przedstawiona. Poza tym pada stwierdzenie o typowych elementach takiej architektury, co rodzi pytanie czy ten system został w ogóle zaimplementowany czy to tylko spekulacje dotyczące tego, jakie moduły mógłby zawierać, gdyby powstał (to sugeruje styl pisania).

U31. Modne stwierdzenia w załączniku A o redukcji narzutów obliczeniowych i zużycia energii (punkt 6, załącznika) są słuszne w dzisiejszych czasach, ale nie udowodnione przez Autora w trakcie rozprawy.

Powyższe uwagi powinny stać się przyczynkiem do szerszej dyskusji nad kształtem rozprawy podczas jej obrony.

8. Przydatność rozprawy dla nauk technicznych i przemysłu

Uważam, że przedłożona rozprawa doktorska Pana Szymona Niewiadomskiego wpisuje się w bieżące problemy informatyki. Opracowanie rozwiązań z dziedziny sztucznej inteligencji w obszarze poprawy jakości danych i struktury baz zarządzania konfiguracją CMDB pozwoliło Autorowi na osiągnięcie zakładanych celów rozprawy przy zachowaniu satysfakcjonującej efektywności. Czy przekłada się to bezpośrednio na tworzenie lepszych rozwiązań w tym obszarze w stosunku do istniejących rozwiązań opublikowanych w światowej literaturze? Trudno powiedzieć. Wydaje się, że zaproponowane rozwiązania rozszerzają w jakimś sensie spektrum istniejących metod stosowanych w obszarze dobrze działających i spójnych baz CMDB.

Reasumując, wyniki osiągnięte przez Pana Szymona Niewiadomskiego w trakcie realizowanych przez niego badań pozwalają potwierdzić, iż opracowane metody mogą być przydatne do rozwiązania zidentyfikowanych problemów przemysłowych. Analiza treści rozprawy wskazuje, że spełnia ona formalne wymagania stawiane rozprawom doktorskim w rozumieniu obowiązujących przepisów. Wnoszę zatem o przyjęcie rozprawy doktorskiej i dopuszczenie jej do publicznej obrony.

Dariusz Mrozek

prof. dr hab. inż. Dariusz Mrozek
Katedra Informatyki Stosowanej
Politechnika Śląska w Gliwicach