Białystok, 21 października 2023 r.

dr hab. Zenon A. Sosnowski, prof. PB
Wydział Informatyki, Politechnika Białostocka
15-351 Białystok, ul. Wiejska 45A
e-mail: z.sosnowski@pb.edu.pl

RECENZJA

rozprawy doktorskiej dla
Rady Naukowej Instytutu Badań Systemowych PAN

Tytuł rozprawy: Automatyczna klasyfikacja dokumentów tekstowych w zarządzaniu aktami spraw na użytek e-administracji

Autor rozprawy: mgr Marek Gajewski

1. Wstęp

Recenzja opracowana została na podstawie umowy z IBS PAN reprezentowanym przez Zastępcę Dyrektora ds. Naukowych dr. hab. inż. Jana W. Owsińskiego.
Przedstawiona do recenzji rozprawa składa się z ośmiu rozdziałów, bibliografii (94 pozycje, w tym 10 współautorstwa doktoranta). Łączna objętość rozprawy wynosi 93 strony.

2. Merytoryczna zawartość rozprawy doktorskiej

Tematyka rozprawy doktorskiej mgr. Marka Gajewskiego związana jest z metodami klasyfikacji dokumentów tekstowych. Jest to zagadnienie ciszące się od kilku lat niesłabnącym zainteresowaniem, głownie w kontekście rozwoju, na całym świecie i w Polsce, koncepcji e-administracji.
Układ pracy jest poprawny. Praca została podzielona na osiem rozdziałów, które kolejno wprowadzają do dziedziny wyszukiwania informacji tekstowej i jej narzędzi. Pierwsze cztery rozdziały wprowadzają podstawowe pojęcia i definicje potrzebne w niniejszej rozprawie doktorskiej. W szczególności, w rozdziale pierwszym jasno określono cel i motywację pracy oraz przedstawiono hipotezę badawczą. Rozdział

drugi przedstawia dziedzinę wyszukiwania informacji tekstowej w ujęciu klasycznym. Rozdział trzeci poświęcono przedstawieniu pokrótce tych elementów teorii zbiorów rozmytych, które są istotne dla zaproponowanego w pracy rozwiązania, w szczególności pojęcie zawierania się zbiorów rozmytych. Rozdział czwarty poświęcony jest krótkiemu przedstawieniu zadania wykrywania i śledzenia wątku (ang. *Topic Detection and Tracking, TDT*), jako zadania rozważanego w literaturze i najbliższego zadaniu postawionemu w rozprawie.

W rozdziale piątym przedstawiona jest nowa autorska metoda, który pozwoli konstruować, na podstawie próby uczącej dokumentów, klasyfikator efektywnie przypisujący dokumenty do odpowiednich kategorii (klas) i spraw (sekwencji dokumentów). Przyjęta reprezentacja danych oparta jest na klasycznym modelu wektorowym. Zaproponowano reprezentację: dokumentu, sprawy (sekwencji dokumentów) oraz kategorii. Proponowana reprezentacja dokumentu używa ograniczonej liczby słów kluczowych, odpowiadających współrzędnym o największych wartościach w wektorze wszystkich słów kluczowych. Sekwencje dokumentów, tworzących sprawy, reprezentowane są przez połączenie wektorów reprezentujących kolejne składowe dokumenty. Przyjęta w proponowanym podejściu reprezentacja kategorii, przybiera postać średniej wszystkich wektorów reprezentujących poszczególne dokumenty aktualnie przypisane do określonej kategorii. W zaproponowanym przez Doktoranta bazowym algorytmie, klasyfikacja dokumentu odbywa się na podstawie oceny jego przynależności do poszczególnych spraw i kategorii z nimi związanych. Dla każdego nowego dokumentu oraz każdej sprawy obliczane są dwa stopnie podobieństwa, odpowiednio: dotyczący danej sprawy i drugi dotyczący kategorii, do której sprawa ta należy. Ostatecznie dokument przypisywany jest do tej sprawy, dla której wyliczona ważona suma podobieństw jest największa. Kluczowym w zaproponowanym w rozprawie podejściu jest przyjęty sposób obliczania stopni podobieństwa odwołujący się do pojęcia zawierania się zbiorów rozmytych. Algorytm bazowy został następnie rozszerzony na przypadek hierarchicznie zorganizowanego zbioru kategorii (klas). Przyjęto, że kategorie, do których należy przypisać dokumenty, stanowią liście pewnej struktury hierarchicznej (drzewiastej). W zaproponowanym przez Doktoranta podejściu uwzględnienie istnienia hierarchii kategorii czy całej sekwencji dokumentów w sprawie sprowadza się do różnicowania ważności słów kluczowych reprezentujących kategorie i sprawy. Sama klasyfikacja dokumentów odbywa się na tych samych zasadach co dla zadania podstawowego, a więc zasadniczym pojęciem

jest nadal stopień zawierania się zbiorów rozmytych. Wymaga to jednak uwzględnienia ważności słów kluczowych przy obliczaniu tego stopnia. W tym celu zaproponowano dwa rozszerzenia pojęcia zawierania się zbiorów rozmytych z uwzględnieniem stopnia ważności elementów przestrzeni rozważań. Pierwsze rozszerzenie oparte jest na wskaźniku Kosko, który w naturalny sposób uwzględnienia stopnia ważności elementów. Ważność słów kluczowych utożsamiana jest ze zbiorem rozmytym w przestrzeni wszystkich słów kluczowych, a do jej uwzględnienia wykorzystano rachunek kwantyfikatorów lingwistycznych Zadeha. W drugim rozszerzeniu wskaźnika zawierania się Kosko w celu uwzględnienia stopni ważności elementów jest interpretacja tego wskaźnika w terminach prawdopodobieństwa warunkowego zdarzenia rozmytego. Przeprowadzono szczegółową analizę własności proponowanych wskaźników zawierania się.

Przeprowadzone eksperymenty obliczeniowe oraz analiza uzyskanych wyników stanowią treść rozdziału szóstego. Opracowanie teoretyczne proponowanego w rozdziale piątym rozwiązania zostało przetestowane w praktyce. Badania przeprowadzone zostały na dwóch zbiorach danych. Przeprowadzone eksperymenty obliczeniowe miały na celu porównanie zaproponowanej metody z popularnym klasyfikatorem k-nn oraz zweryfikowanie efektywności różnych parametryzacji tej metody, to jest efektów zastosowania różnych wskaźników zawierania się zbiorów rozmytych oraz wag obydwu kryteriów, to jest dopasowania do kategorii i dopasowania do sprawy. W pierwszej kolejności przeprowadzono szereg wstępnych obliczeń w celu wyboru konfiguracji parametrów do szczegółowych eksperymentalnych badań porównawczych. W kolejnym eksperymencie porównano zaproponowaną metodę z dwoma wariantami algorytmu typu *k*-nn. Zaproponowana przez Autora klasa algorytmów wykazała wyraźną przewagę nad zastosowanymi wersjami algorytmu 1-nn.

Rozdziały siódmy pełni w rozprawie rolę przeglądu literatury. Przedstawiono są w nim inne podejścia do rozwiązania zadania stanowiącego tezę rozprawy, rozważane przez Autora na wcześniejszych etapach przygotowania rozprawy doktorskiej. Obejmowały one algorytm k najbliższych sąsiadów (*k*-nn), maszyny wektorów podpierających (ang. support vector machine, SVM) oraz metody oparte na modelowaniu sekwencji t,j. ukryty model Markowa (ang. Hidden Markov Model; HMM) i zastosowanie algorytmu eksploracji sekwencji.

W ostatnim, ósmym rozdziale, Doktorant podsumował przeprowadzone badania zaznaczając potrzebę dalszych eksperymentów na różnych zbiorach dokumentów, aby możliwe było wyciągnięcie dalej idących wniosków.

3. Ocena rozprawy w aspekcie merytorycznym

Rozprawa ma zarówno charakter doświadczalny jak i teoretyczny. Doktorant głównie koncentruje się w niej na problemie kategoryzacji dokumentów tekstowych w oparciu o zaproponowany przez niego algorytm. Ostatecznym celem algorytmu jest klasyfikacja, w której każdy (tekstowy) dokument, oprócz przynależności do właściwej kategorii, należy do określonej sekwencji dokumentów, rozumianej jako sprawa, zawierającej dokumenty z tej samej kategorii.

Pozytywnie oceniam przyjętą przez Autora metodologię badań. Autor dokonuje krytycznej analizy stanu badań w dziedzinie klasyfikacji dokumentu zgodnie z jego szeroko rozumianą tematyką, z drugiej zaś strony, zgodnie z pewnym procesem, sprawą, której on dotyczy. W oparciu o tę analizę proponuje własne rozwiązanie, a następnie przeprowadza szereg badań eksperymentalnych. Przeprowadzone eksperymenty stanowią istotny element rozprawy. Doktorant bada zaproponowane w rozprawie algorytmy, porównując je z propozycjami z literatury. Uzyskane eksperymentalnie wyniki potwierdzają skuteczność proponowanych klas algorytmów.

Oryginalny dorobek rozprawy zawarty jest w rozdziałach 4-7. Do dorobku tego należy zaliczyć:

a) Opracowanie nowatorskiego rozwiązania problemu kategoryzacji dokumentów tekstowych, w którym każdy (tekstowy) dokument, oprócz przynależności do właściwej kategorii, należy do określonej sekwencji dokumentów, rozumianej jako sprawa, zawierającej dokumenty z tej samej kategorii.

b) Projekt i realizację eksperymentów obliczeniowych pozwalających potwierdzić tezę rozprawy mówiącą o tym, że „można skonstruować efektywny algorytm dla przedstawionego wyżej zadania, przyjmując bardzo uproszczoną reprezentację dokumentów w postaci zbiorów słów kluczowych o niewielkiej liczności oraz stosując pojęcie zawierania się zbiorów (rozmytych) jako podstawę do zaklasyfikowania dokumentu do sprawy."

c) Największą wartość merytoryczną w pracy jest rozdział piaty p.t. „Metoda rozwiązania". Proponowany algorytm wykorzystuje pojęcie zawierania się zbiorów rozmytych z uwzględnieniem stopni ważności poszczególnych

elementów przestrzeni rozważań. W rozprawie zaproponowano rozszerzenie poszczególnych wskaźników zawierania się zbiorów z uwzględnieniem dodatkowych stopni ważności poszczególnych elementów.

Rozprawa jako całość nie ma istotnych wad. Wśród uchybień bądź słabszych stron można wymienić następujące:

- Moim zdaniem dwa zbiory danych użyte w eksperymentach obliczeniowych nie są zbytnio zróżnicowane Liczba kategorii i spraw to podobne rzędy wielkości. Ciekawe byłyby badania, w których zbiory danych byłyby bardziej zróżnicowane. Warto byłoby wówczas sprawdzić podatność proponowanego algorytmu na taki zróżnicowany charakter danych.
- Praca dotyczy dyscypliny informatyka techniczna. Spodziewałem się nieco większego wkładu Doktoranta w zagadnienia związane z analizą proponowanych algorytmów w zakresie np. złożoności obliczeniowej. Proponowana metoda jest oceniona w kategorii jakości klasyfikacji, ale jej ocena nie musi ograniczać się tylko do tego aspektu. Innymi potencjalnie interesującymi kategoriami oceny tego typu metod mogą być np. złożoność obliczeniowa/pamięciowa.
- W rozdziale 6.4.3 brakuje porównania jak kształtuje się skuteczność klasyfikacji wraz ze wzrostem liczby toczących się ("otwartych") spraw użytym do badań w rozdziale poprzednim algorytmem k-nn.
- W recenzowanej rozprawie zabrakło mi szerszej dyskusji o ograniczeniach dla zaproponowanej klasy algorytmów. Potencjalny użytkownik dysponujący konkretnymi danymi chciałby wiedzieć, czy możliwe jest zastosowanie proponowanych algorytmów za pomocą posiadanej infrastruktury.
- Wyniki poprzednich badań przedstawione w rozdziale 7 ma bazie wcześniejszych publikacji Autora nie są porównywalne, ze względu na wykorzystane tylko znacznie mniejszego podzbioru dokumentów.
- W pracy zdefiniowano wiele pojęć. W mojej ocenie przydatny byłby osobny wykaz wszystkich symboli i oznaczeń.
- Zauważyłem następujące drobne błędy redakcyjne:
  - ✓ Str. 20: warto dodać, że *max* to najmniejsza norma *s*
  - ✓ Str. 23, wzór (69): „$x \geq 0.8$" -> „$u \geq 0.8$"
  - ✓ Str. 32, wzór (87): Nie jest jasne czy w sekwencji są brane te same słowa kluczowe.

- ✓ Str. 33, wzór (91): Nie jest wyjaśnione, dlaczego dodanie dokumentu d_2 nie zostało wykonane zgodnie ze wzorem (89).

Przegląd literatury dotyczącej zagadnień, którym poświęcona jest rozprawa jak i sformułowane przez Doktoranta wnioski z niego wynikające są w zasadzie wystarczające. Niewątpliwie są one dowodem dużej wiedzy Autora w danej dyscyplinie naukowej.

Oceniając całościowy dorobek rozprawy stwierdzam, że Doktorant wykazał umiejętność poprawnego i przekonywującego przedstawienia uzyskanych przez siebie wyników dobrze je dokumentując. Przygotowując rozprawę Doktorant wykazał się dużą starannością oraz pracowitością. Praca jest przydatna dla nauk inżynieryjno-technicznych.

4. Wniosek końcowy

Wszystkie moje uwagi krytyczne i dyskusyjne w żadnym stopniu nie wpływają na jednoznacznie pozytywną ocenę recenzowanej pracy. Stwierdzam, że praca „Automatyczna klasyfikacja dokumentów tekstowych w zarządzaniu aktami spraw na użytek e-administracji" spełnia wymagania stawiane rozprawom doktorskim zgodnie z Ustawą z dnia 14 marca 2003 roku o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki (Dz. U. z 2017 r. poz. 1789, z późn. zm.). W mojej ocenie rozprawa zawiera oryginalne rozwiązanie problemu. Doktorant osiągnął stawiany cel, wykazując się niezbędną ogólną wiedzą teoretyczną w dyscyplinie informatyka techniczna i telekomunikacja oraz umiejętnościami do samodzielnego rozwiązywania problemów naukowo-technicznych W związku z powyższym wnioskuję o przyjęcie recenzowanej rozprawy jako rozprawy doktorskiej i dopuszczenie mgr. Marka Gajewskiego do jej publicznej obrony.

Zenon A. Sosnowski

/Zenon A. Sosnowski/