Dr hab. Hung Son Nguyen
Wydział Matematyki, Informatyki i Mechaniki
Uniwersytet Warszawski
email: son@mimuw.edu.pl

Warszawa, 20/4/2019

Recenzja rozprawy doktorskiej

Tytuł:

Metody semantycznej kategoryzacji w zadaniach analizy dokumentów tekstowych

Autor rozprawy: **Mgr Piotr Borkowski**

Promotor: **prof. dr hab. Inż. Mieczysław A. Kłopotek**

Promotor pomocniczy: **dr Krzysztof Ciesielski**

Rozprawa doktorska została wykonana w:

Instytucie Podstaw Informatyki
Polskiej Akademii Nauk

Recenzowana rozprawa opisuje pewną metodę kategoryzacji dokumentów w oparciu o taksonomię pojęć, manualnie utworzoną przez społeczność internetową. Autor chciałby pokazać, że takie taksonomie mogą być traktowane jako bazy wiedzy dla algorytmów semantycznej analizy tekstów. Proponowane rozwiązanie powinno być proste w implementacji i niewymagające przygotowywania odpowiednich etykietowanych zbiorów treningowych. Autor rozprawy proponował algorytm kategoryzacji semantycznej oraz algorytm klasyfikacji semantycznej oraz przeprowadził szereg eksperymentów nad rzeczywistymi danymi pochodzącymi z publicznych źródeł.

1. ZAKRES ROZPRAWY

Problematyka rozprawy doktorskiej mgr. Piotra Borkowskiego mieści się w dziedzinie text mining, w szczególności semantycznej analizy dokumentów tekstowych. Jest to dziedzina, która łączy ze sobą różne aspekty informatyki, zarówno teoretyczne jak i praktyczne, algorytmiczne, czy też obliczeniowe.

Pan mgr Piotr Borkowski podjął się próby zbadania zagadnienia semantycznej kategoryzacji dokumentów tekstowych. Pomysł polega na wykorzystaniu istniejących zasobów tekstowych w postaci taksonomii jako źródło wiedzy do konstrukcji kategoryzatora (tzn. algorytmu kategoryzacji). Jest to alternatywne podejście i wspierająca metoda dla istniejących algorytmów klasyfikacji z tą różnicą, że w proponowanym rozwiązaniu, nie jest konieczne przygotowanie odpowiednich zbiorów treningowych. Autor rozprawy zbadał również możliwe zastosowania algorytmu semantycznej kategoryzacji w zadaniach klasyfikacji tekstów. Takie narzędzia są coraz bardziej potrzebne, zwłaszcza w obecnej erze globalnej cyfryzacji i mamy do czynienia z coraz większymi repozytoriami dokumentów tekstowych.

Mgr Piotr Borkowski zbadał również skuteczności proponowanych metod zostały realizowane na danych pochodzących z największymi taksonomiami takimi jak Wikipedia i MESH oraz sprawdzone na zbiorach danych benchmarkowych zarówno w języku polskim jak i angielskim.

Zatem podjętą problematykę badawczą należy uznać za ważną i aktualną, a jej wybór za słuszny. Postawiona teza też jest bardzo interesująca i jest godnym tematem rozprawy doktorskiej.

2. STRUKTURA ROZPRAWY

Recenzowana praca zawiera 144 strony i składa się 10 rozdziałów wraz ze streszczeniem i bibliografią.

W pierwszym rozdziale Pan mgr Piotr Borkowski przedstawił motywację, określił problem badawczy w rozprawie i plan jej realizacji. Autor ustalił też metodologię badań, czyli metodę przeprowadzania eksperymentów w celu weryfikacji skuteczności proponowanych rozwiązań.

Drugi rozdział zawiera przegląd literatur związanych z tematyką rozprawy. Autor dokonał klasyfikacji technik kategoryzacji dokumentów oraz klasyfikacji dokumentów i przedstawił odpowiednie literatury dla każdej grupy metod. Koncentrował on głównie na przeglądzie technik eksploracji dokumentów tekstowych z wykorzystaniem zasobów semantycznych. Następnie, Pan mgr Piotr Borkowski omówił stan wiedzy na temat technik radzenia sobie z problemem wieloznaczności słów zarówno dla języka angielskiego jak i polskiego. Na końcu tego rozdziału, Autor rozprawy przedstawił dość szczegółowo projekt Wikipedia-Milner, który ma wiele wspólnych cech i celów jak w recenzowanej rozprawie i próbował podkreślić różnice między projektami. Autor uzupełnił też informacje o innych projektach wykorzystujących Wikipedię.

Trzeci rozdział może być uważany za kluczowy dla rozprawy. W tym rozdziale, Autor zaproponował algorytm kategoryzacji dla dokumentów tekstowych bazujący na taksonomii pojęć. Jest to algorytm korzystający z większości istniejących taksonomii pojęć i założenia dotyczące możliwych taksonomii zostały również omówione w tym rozdziale. Proponowany algorytm składa się z takich procesów jak przygotowywanie danych, mapowanie (przekształcanie) termów na koncepty, ujednoznacznienie wieloznacznych mapowań z

wykorzystaniem autorskiej miary podobieństwa semantycznego oraz mapowanie konceptów na listę kategorii z wagami. Te procesy zostały szczegółowo opisane w rozdziale trzecim.

W czwartym rozdziale Autor przedstawił różne miary semantycznego podobieństwa i wyróżnił dwa rodzaje miar, tj. IC (Information Content) i MSCA (Most Specific Common Abstraction). Za pomocą tychże miar autor zdefiniował dwie miary podobieństwa, jedna mierzy podobieństwo między kategoriami, a druga mierzy podobieństwo między konceptami.

Rozdziały piąty i szósty zawierają szczegółowe opisy procesów przygotowania danych z Wikipedii i przekształcania ich do odpowiedniego formatu taksonomii pojęć. Tak przygotowywana taksonomia pojęć z Wikipedii stanowi bazę dla działania algorytmu kategoryzacji polskojęzycznych dokumentów tekstowych.

Oceny jakości działania algorytmu kategoryzacji na rzeczywistych danych zostały przedstawione w rozdziale siódmym. Autor opisał 4 zbiory dokumentów biorących udział w eksperymentach, tj. (1) zbiór DMOZ (2804 dokumenty z 34 kategorii); (2) zbiór kopalniawiedzy.pl (13099 dokumentów); (3) zbiór dokumentów z 6 wybranych polskich domen (8503 dokumentów) oraz zbiór dokumentów medycznych w języku angielskim BioASQ. Do kategoryzacji dokumentów w języku polskim autor używał taksonomii utworzonej z struktur dokumentów z Wikipedii, zaś do kategoryzacji dokumentów w języku angielskim używał taksonomii MeSH. W rozdziałach 7.2 i 7.3 Autor opisał techniki oceniania algorytmów kategoryzacji oraz porównywania wyników. W kolejnych rozdziałach Autor przedstawił pomiarowe wyniki eksperymentów nad różnymi modułami ujednoznaczniania, nad skutecznością procesu kategoryzacji i nad wpływami różnych technik kategoryzacji. Na koniec autor przedstawił również przykład działania algorytmu kategoryzacji dla dokumentów medycznych w języku angielskim.

Rozdział ósmy opisuje dwie techniki klasyfikacji semantycznej. Pierwsza metoda nazywa się SemCla i bazuje się na kategoryjnej reprezentacji dokumentu, tj. rozszerzonym wektorze kategorii wraz z wagami. Druga metoda jest heterogenicznym komitetem klasyfikatorów zawierającym metodę Naive Bayes, Wide-Margin Winnow, L-LDA oraz algorytm kategoryzujący SemCat.

W rozdziale dziewiątym Autor przedstawił wyniki eksperymentów zbiorach dokumentów w języku polskim opisanych wcześniej w rozdziale siódmym. Autor rozprawy korzystał z testu Manna-Whitneya-Wilcoxona do porównywania różnych technik klasyfikacji. Wyniki eksperymentów pokazały, że proponowany algorytm kategoryzacji jest skutecznym narzędziem do konstrukcji efektywnych klasyfikatorów semantycznych.

Rozdział dziesiąty przedstawia wnioski i kierunki dalszych badań.

Bibliografia zawiera spis ponad 100 artykułów naukowych związanych z tematyką rozprawy.

3. OCENA ZAWARTOŚCI ROZPRAWY

Rozprawa doktorska mgr. Piotra Borkowskiego przedstawia trzy główne rezultaty:

I. Algorytm semantycznej kategoryzacji dokumentów tekstowych wykorzystujący taksonomię pojęć jako bazę wiedzy.

Ten algorytm działa jako funkcja, która przyjmuje na wejściu dokumenty tekstowe i zwraca na wyjściu listy rankingowe kategorii związanych z tymi dokumentami. Proponowany algorytm, zwany **SemCat**, składa się z trzech głównych transformacji:

a. przekształcenie dokumentu w wektor termów z wagami według schematu tfidf

b. mapowanie wektora termów na zbiór konceptów

c. mapowanie zbioru konceptów na listę kategorii z wagami (rankingami)

Pierwszy etap jest standardowym procesem wstępnego przetwarzania tekstów i jest typowy dla wielu algorytmów w dziedzinie "text mining". Ten etap wymaga znajomości języka i składa się z takich kroków jak wykrywanie języka, czyszczenie, usunięcie stop words, lematyzacja, identyfikacja frazów, i wyznaczanie wag TFIDF dla termów.

W mojej ocenie, najważniejszym pomysłem w recenzowanej rozprawy jest dołożenie przestrzeni konceptów pomiędzy przestrzenią termów a przestrzenią kategorii. W ten sposób proces kategoryzacji jest złożeniem dwóch procesów: mapowanie wektora termów na zbiór konceptów i mapowanie zbioru konceptów na listę kategorii z wagami. Różnica między konceptami a kategoriami polega, m.i. na tym, że kategorie są pojęciami z zadanej taksonomii pojęć i są ze sobą związane w pewnej strukturze taksonomicznej, czyli graf skierowany, acykliczny z tylko jedną najbardziej ogólną kategorią, wówczas gdy koncepty mogą być w pewnej częściowej relacji lub nawet w pustej relacji. Można to porównywać z podejściem warstwowego uczenia się (layered learning). Mapowanie wektorów termów na koncepty odbywa się za pomocą opracowanej przez autora miary podobieństwa oraz algorytmu ujednoznaczniania zbioru konceptów. Algorytm mapowania kategorii polega na wyznaczniu ważonej listy kategorii związanych z danym zbiorem konceptów, a następnie przekształceniu takiej listy w zbiór wynikowy.

II. Zastosowanie wyżej wymienionego algorytmu kategoryzacji do konstrukcji semantycznego klasyfikatora dla dokumentów tekstowych

Autor rozprawy proponował pewien algorytm klasyfikacji dokumentów w oparciu o algorytm kategoryzacji. Pomysł polega na użyciu wektorów kategorii z wagami zamiast wektorów słów do reprezentowania zarówno dokumentów ze zbioru treningowego jak i dokumentów ze zbioru testowego. Ta nowa reprezentacja może być używana do klasyfikacji za pomocą dowolnego algorytmu klasyfikacji tekstów. Autor rozprawy zaproponował specjalną metodę klasyfikacji zwaną **SemCla**, która jest heterogeniczny komitetem klasyfikatorów złożonych z Naive Bayes, Wide-Margin Winnow, L-LDA i wspomnianego wcześniej algorytmu kategoryzacji SemCat.

III. Eksperymenty na danych rzeczywistych w celu weryfikacji jakości proponowanych rozwiązań.

Autor rozprawy przeprowadził wiele eksperymentów dla obu proponowanych algorytmów: SemCat i SemCla.

Autor rozprawy przeprowadził eksperymenty nad algorytmem SemCat dla dwóch wersji językowych: polskiej i angielskiej. Dla języka polskiego autor wyekstrahował strukturę taksonomiczną z Wikipedii i korzystał z niej jako baza wiedzy, a następnie uruchomił to narzędzie na czterech zbiorach danych: DMOZ, kopalniawiedzy.pl oraz zbiór dokumentów z 6 wybranych polskich domen.

Eksperymenty z językiem angielskim zostały przeprowadzone za pomocą taksonomii MESH na zbiorze dokumentów z konkursu BioASQ zostały przedstawione w postaci przykładów działania algorytmu dla niektórych dokumentów (bez szczegółowych analiz tak jak dla dokumentów w języku polskim).

Autor rozprawy używał trzech miar: miara poprawności Lin, miara poprawności shortest path oraz binarnej miary poprawności do badania nad skuteczności algorytmów kategoryzacji i klasyfikacji oraz wpływu różnych parametrów na jakości tych algorytmów.

Skuteczność algorytmu klasyfikacji została zbadana za pomocą eksperymentów nad dokumentami w języku polskim. Do porównywania jakości dwóch klasyfikatorów, Autor używał miary Manna-Whitneya-Wilcoxona. Eksperymenty pokazują, że proponowana metoda klasyfikacji tekstów za pomocą heterogenicznego komitetu klasyfikatorów SemCla w wielu przypadkach okazała się lepsza niż inne standardowe metody, które zostały wybrane przez Autora do analizy.

4. UWAGI KRYTYCZNE

Pozytywnie oceniam wyniki osiągnięte przez doktoranta i opisane w rozprawie. Ale mam też pewne krytyczne uwagi do przedłożonej wersji rozprawy. Niżej przetoczę najważniejsze z nich:

1. Uwagi dotyczące zawartość rozprawy: uważam, że rozprawa powinna zawierać pewne dodatkowe wyniki badań i analizy, aby móc ją traktować jako kompletne dzieło dotyczące tematyki rozprawy:
   - W prawdzie, Pan mgr Piotr Borkowski przedstawił dwie autorskie metody semantycznej analizy dokumentów tekstowych, ale pierwsza metoda, zwana semantyczna kategoryzacja dokumentów, można traktować jako proces odkrywania cech (ang. feature extraction) dla metody klasyfikacji dokumentów. Jedyną nowością w przedstawionej metodzie klasyfikacji dokumentów jest właśnie ta wspomniana metoda kategoryzacji. Dlatego czuję pewien niedosyt, bo metoda kategoryzacji może służyć również metodom grupowania semantycznego lub metodom wyszukiwania semantycznego. Pod koniec rozprawy, autor wspomniał o tym, że powyższe metody

sprawdziły się w wyszukiwarce NEKST, ale nie podał szczegółowych informacji o ich skuteczności. W tym sensie, tematyka rozprawy nie została kompleksowo zbadana.

- Autor rozprawy koncentrował się na badaniu skuteczności opracowanych metod za pomocą zbiorów dokumentów w języku polskim. Z jednej strony, to jest bardzo cenne, że metody zostały opracowane i optymalizowane dla języka polskiego. Ale z drugiej strony, brak testów jakościowych na dokumentach w języku angielskim nie pozwala na porównywanie proponowanych metod z najlepszymi metodami na świecie.
- Przy analizie skuteczności proponowanych metod autor nie zbadał złożoności obliczeniowych. Brak takiej analizy spowodowała to, że ocena możliwości zastosowania tych rozwiązań w praktycznych zastosowaniach stała się niemożliwa.

2. Uwagi dotyczące edycji rozprawy: w mojej ocenie, proponowane metody w rozprawie nie były opisywane w sposób kompletny i precyzyjny:
   - Autor nie używał wyróżnionych środowisk do definiowania ważnych pojęć i opisywania algorytmów. Wszystkie algorytmy i ich komponenty zostały przedstawione jedynie w postaci pseudokodu. Brakuje opisu zarówno danych wejściowych jak i wyjściowych. Brakuje również opisu struktur danych potrzebnych do implementacji proponowanych algorytmów i ich składowych komponentów. Brak wyróżnionego środowiska dla algorytmów i definicji pojęć istotnie zmniejszył czytelność rozprawy.
   - Brakuje również przykładów ilustrujących różne pojęcia i metod. W rozprawie autor używał dość podobnych określeń do oznaczenia różnych pojęć, np. kategorie, koncepty, etykiety czy zbiór tytułowy.
   - Są pewne niedociągnięcia edytorskie. Np.
     - W ósmym rozdziale Autor opisał metodę komitetu klasyfikatorów i nazwał ją SemCom, a w rozdziale dziewiątym zbadał skuteczności metody ale nigdy nie używał tej nazwy.
     - Na stronie 113 algorytm SimCat był omyłkowo nazwany „klasyfikatorem" a nie kategoryzatorem jak w innych miejscach rozprawy.
     - W opisie algorytmu na stronie 45 pojawiła się funkcja $sim(k_i, l)$, która nie została wcześniej ani definiowana, ani komentowana. Przez domniemanie można zgadywać, że ona może być jedną z funkcji opisanych w następnym rozdziale.
     - W algorytmie SemCat dokumenty są reprezentowane przez listy termów z wagami (zamiast pojedyńczych słów), podczas gdy w algorytmie klasyfikacji Naive Bayes, autor znów używał listy słów. Czy ten wybór był świadomy, czy tylko pomyłka?
     - Zdarzają też literówki w niektórych miejscach. Np. na stronie 41, linia 8: powinno być „zawierających" zamiast „zawierająca"

3. Publikacja: Spis publikacji nie był posortowany według nazwisk autorów, przez co wyszukiwanie informacji było bardzo utrudnione. Na przykład, przy przeszukiwaniu nazwiska Doktoranta zauważyłem, że był on autorem czterech publikacji (możliwe, że jest

więcej, ale nie byłem w stanie ich wykryć). Dwie publikacje ([55], [66]) były artykułami opublikowanymi na międzynarodowych konferencjach (SIIS 2011 i ISMIS 2014), publikacja [106] była po polsku, a publikacja [107] jest archiwalnym artykułem. W żadnym z tych publikacji Doktorant nie był samodzielnym autorem. Zatem oceniam dorobek naukowy jako słaby.

5. UWAGI KOŃCOWE

Reasumując można stwierdzić, że recenzowana praca doktorska zawiera interesujące wyniki. Rozprawa stanowi samodzielne rozwiązanie przez Doktoranta problemu naukowego, gdyż Autor wykazał się umiejętnością identyfikacji problemów badawczych, formułowania celu badań, pracy nad badaniami literaturowymi w zakresie analizowanych problemów, konstruowania i doboru metod badawczych, przeprowadzenia badań, wnioskowania i prezentacji wyników,

Po lekturze rozprawy można również stwierdzić, że Autor posiada ogólną wiedzę w dyscyplinie informatyki. Mimo, że w recenzji zostały wskazane pewne krytyczne uwagi oraz zgłoszone pewne zastrzeżenia, praca stanowi jednak ciekawy, oryginalny przykład zastosowania teorii analizy danych w praktyce oraz stanowić cenny materiał dla praktycznych zastosowań.

Uwzględniając wszelkie uwagi - zarówno aprobujące, jak i krytyczne oraz mając świadomość istnienia w przedstawionej do recenzji pracy pewnych kwestii dyskusyjnych, stwierdzam, że praca mgr. Piotra Borkowskiego spełnia wymogi stawiane pracom doktorskim. Wnoszę zatem o dopuszczenie przedłożonej mi do recenzji rozprawy do publicznej obrony.

Hung Son Nguyen