dr hab. inż. Piotr Gawrysiak, prof. PW
Instytut Informatyki
Wydział Elektroniki i Technik Informacyjnych
Politechnika Warszawska

Warszawa, 29 marca 2019

RECENZJA ROZPRAWY DOKTORSKIEJ MGR. PIOTRA BORKOWSKIEGO DLA RADY NAUKOWEJ INSTYTUTU PODSTAW INFORMATYKI POLSKIEJ AKADEMII NAUK

Tytuł rozprawy: Metody semantycznej kategoryzacji w zadaniach analizy dokumentów tekstowych.

1. Jaki jest problem naukowy (teza) rozprawy i czy został on trafnie i jasno sformułowany?

Recenzowana rozprawa poświęcona jest problematyce analizy języka naturalnego, w tym w szczególności problemowi określania tematyki dokumentów, zawierających treść w języku naturalnym. Celem przeprowadzonych przez Autora prac było przede wszystkim opracowanie algorytmów, pozwalających w sposób automatyczny, bez procesu uczenia nadzorowanego, na przypisanie do dokumentu tekstowego elementu zadanej ontologii, zgodnie z semantyką tego dokumentu. Algorytmy te, nazywane przez Autora algorytmami kategoryzacji dokumentów[1], mogą być następnie wykorzystywane samodzielnie, bądź też jako element systemów automatycznej klasyfikacji dokumentów.

Teza sformułowana w rozprawie nie zawiera założeń dotyczących struktury taksonomii, która ma być źródłem etykiet nadawanych przetwarzanym dokumentom. Oczywistym jednak jest, iż celem Autora było umożliwienie wykorzystania dużych taksonomii, zawierających tysiące i więcej elementów (patrz strona 27 rozprawy). Innymi słowy takich, których wykorzystanie bądź to w procesie ręcznej obróbki zbiorów dokumentów, bądź też w systemach wykorzystujących algorytmy uczenia maszynowego (i tym samym także wymagających ręcznego tworzenia zbiorów uczących) byłoby kłopotliwe lub wręcz niemożliwe.

1 Terminologia użyta przez Autora może być dyskusyjna, jako że słowa „klasyfikacja” oraz „kategoryzacja” nie wydają się być wystarczająco precyzyjne i bywają często mylone, nie tylko w mowie potocznej. Dotyczy to także ich angielskich odpowiedników „categorization” i „classification” - por. np. *Jacob, Elin K.. “Classification and Categorization: A Difference that Makes a Difference.” Library Trends 52 (2004): 515-540.* W przypadku recenzowanej rozprawy być może właściwszym byłoby zastosowanie określenia „etykietowanie” zamiast „kategoryzacja”.

2. Czy Autor rozwiązał postawiony problem i czy użył do tego właściwych metod dowodząc, że posiadł umiejętności związane z metodyką i metodologią prowadzenia badań naukowych?

Większość współczesnych systemów semantycznej analizy języka naturalnego wykorzystuje algorytmy uczenia maszynowego, mające swoje korzenie w metodach opracowanych w początkach kształtowania się dyscypliny eksploracji danych (*data mining*). Ogólniej, dotyczy to także systemów analizujących znaczenie innych – niż język naturalny – treści, takich jak obrazy, czy też dźwięk. Systemy te, dzięki coraz bardziej efektywnym algorytmom (należy tu przede wszystkim wymienić, zyskujące w ostatnich latach szczególną popularność, tzw. głębokie sieci neuronowe) osiągają spektakularne efekty głównie w zadaniach klasyfikacji treści. Warunkiem koniecznym ich użyteczności jest jednak dostępność odpowiednio dużej liczby uprzednio sklasyfikowanych przykładów - to zaś może nie być trywialne. Nierzadko o wartości (w tym także rynkowej) przedsiębiorstw, oferujących systemy sztucznej inteligencji, decyduje obecnie nie tyle sama jakość używanych algorytmów i metod, co wielkość zgromadzonych danych trenujących, dzięki którym działanie tych algorytmów jest w ogóle możliwe.

Autor recenzowanej rozprawy zdecydował się rozwiązać powyższy problem, opracowując algorytmy nie wymagające dostępności zbiorów uczących. Zaproponowane przezeń metody dokonują „rzutowania" analizowanych dokumentów tekstowych na zadaną hierarchię kategorii, dzięki wykorzystaniu odpowiednio skonstruowanych miar semantycznego podobieństwa dokumentu do opisu kategorii. Oczywiście stworzenie tego rodzaju metod wymagało, poza samym opracowaniem wspomnianych miar, rozwiązania także problemów dotyczących niejednoznaczności elementów hierarchii kategorii czy też odpowiedniego wstępnego przetwarzania zawartości analizowanych dokumentów tekstowych - co zostało przedstawione w rozprawie, lub w cytowanych innych publikacjach Autora.

Autor omawia także metody półautomatycznego tworzenia hierarchii kategorii, które mogą znaleźć zastosowanie w systemach analizy dokumentów tekstowych. Problem ten wydawać się może pobocznym w stosunku do głównej tezy rozprawy, niemniej jednak jego znaczenie nie jest marginalne. Co prawda opracowane przez Autora algorytmy są algorytmami niezależnymi od języka przetwarzanych dokumentów, to jednak domniemywać można, iż jedną z głównych motywacji powstania recenzowanej rozprawy, był brak skutecznych metod analizy tekstu w języku polskim – jedną zaś z przyczyn tego stanu rzeczy jest brak odpowiednich korpusów danych w tym języku. Dzięki metodom przedstawionym w rozprawie możliwe jest tworzenie tego rodzaju korpusów z wykorzystaniem powszechnie dostępnych, tworzonych społecznościowo zbiorów danych, takich jak zasoby polskiej edycji serwisu Wikipedia.

Końcowa część rozprawy zawiera opis badań eksperymentalnych, których celem była ewaluacja porównawcza zaproponowanych przez Autora algorytmów. Wymagało to przygotowania odpowiednich zbiorów danych testowych (pochodzących, poza wspomnianym już serwisem Wikipedia, także z innych ogólnodostępnych źródeł w sieci Internet; ich przystosowanie do celów przeprowadzonych eksperymentów wymagało, nierzadko nietrywialnej, obróbki), oraz opracowania odpowiedniej metodyki badań.

3. Czy tematyka rozprawy jest aktualna lub dostatecznie ważna?

Prace wykonane w ramach przygotowywania recenzowanej rozprawy związane są ściśle z rozwojem wyszukiwarki semantycznej NEKST, opracowywanej w Instytucie Podstaw Informatyki PAN. Wyszukiwarka ta, to właściwie system typu *question answering* (choć określenie to zwykle nie jest stosowane przez twórców NEKST), działający w oparciu o zasoby polskojęzycznej części sieci WWW – serwis bardzo interesujący technologicznie, choć niestety często niedoceniany.

W odróżnieniu od innych tego rodzaju systemów (wymienić można by tu stronę internetową Wolfram Alpha i przede wszystkim wyszukiwarkę internetową Google) jest on dedykowany przetwarzaniu zasobów w języku polskim i udzielaniu odpowiedzi na pytania w nim formułowane. Jest to o tyle istotne, iż jakość działania serwisów uniwersalnych (takich jak wzmiankowane powyżej) jest wciąż znacząco gorsza w przypadku przetwarzania zasobów polskojęzycznych (w porównaniu ze źródłami w języku angielskim). Co więcej, ponieważ firma Google jest *de facto* monopolistą na rynku wyszukiwarek internetowych, tworzenie konkurencyjnych systemów wyszukiwania informacji staje się sprawą kluczową dla zachowania wolności wypowiedzi i swobodnej dystrybucji informacji w sieci Internet.

Niezależnie, abstrahując od kwestii społecznych, czy gospodarczych przedstawionych powyżej, problematyka automatycznego określania tematyki dokumentów w języku naturalnym jest zadaniem, którego w pełni zadowalającego rozwiązania wciąż nie udało się znaleźć – a które jednocześnie ma niebagatelne znaczenie praktyczne.

4. Na czym polega oryginalny dorobek Autora i jakie jest jego znaczenie poznawcze lub przydatność praktyczna dla nauki bądź techniki?

Do głównych osiągnięć Autora, zaprezentowanych w rozprawie, należy zaliczyć przede wszystkim opracowanie nowego algorytmu semantycznej kategoryzacji dokumentów, działającego bez potrzeby uprzedniego przygotowania zbioru uczącego, a wykorzystującego istniejącą hierarchię kategorii. Elementem prac nad skonstruowaniem tego algorytmu było m.in. opracowanie nowych, oraz dostosowanie istniejących miar semantycznego podobieństwa tekstu, oraz stworzenie metod ujednoznaczniania pojęć.

Ponadto, Autor proponuje metody wykorzystania opracowanego algorytmu do budowania mechanizmów klasyfikacji dokumentów tekstowych, w tym w szczególności takich, w których występuje problem luki semantycznej (algorytm SemCla, oraz heterogeniczny komitet klasyfikatorów SemCom). Poza samymi algorytmami analizy semantycznej, Autor proponuje także interesujące metody preprocesingu dokumentów hipertekstowych, pochodzących z zasobów systemów rodziny MediaWiki (w tym w szczególności encyklopedii sieciowej Wikipedia).

Przydatność rozprawy dla nauk technicznych, jest wysoka. Opracowane algorytmy nadają się do bezpośredniego zastosowania zarówno w wyszukiwarkach internetowych jak też systemach bibliotecznych i repozytoriach wiedzy, czy też wszelkiego rodzaju innych bazach pełnotekstowych. Podkreślić przy tym należy to, iż – zgodnie z opisem zamieszczonym w rozprawie – wspomniane algorytmy zostały zaimplementowane w ramach dużego systemu (wyszukiwarka NEKST Instytutu Podstaw Informatyki PAN), zaś ich działanie zweryfikowane praktycznie. Co istotne wykorzystano przy tym zasoby polskojęzycznego Internetu.

5. Czy rozprawa świadczy o dostatecznej wiedzy Autora, wiedzy na zaawansowanym poziomie, o charakterze podstawowym dla dziedziny nauk technicznych oraz o charakterze szczegółowym, odpowiadającej obszarowi prowadzonych badań naukowych?

W rozprawie przedstawiono podstawowe informacje dotyczące etykietowania dokumentów tekstowych (rozdział 2), słowników kontrolowanych (taksonomii i tezaurusów) oraz ontologii (rozdział 3), czy też przetwarzania i reprezentacji dokumentów w języku naturalnym oraz ich automatycznej klasyfikacji (rozdziały 6 i 8).

Struktura logiczna rozprawy, oraz zastosowana metodyka badań, jest poprawna. Pewne wątpliwości budzi jedynie wybór metod ewaluacji jakości działania zaproponowanych algorytmów, oraz pominięcie analizy ich złożoności obliczeniowej.

6. Czy rozprawa obejmuje najnowsze osiągnięcia nauki i świadczy o znajomości współczesnej literatury z dyscypliny naukowej, której dotyczy?

Rozdział drugi rozprawy zawiera przegląd stanu wiedzy dotyczącego problematyki automatycznego etykietowania dokumentów tekstowych, ujednoznaczniania sensów słów oraz przetwarzania zasobów hipertekstowych, w tym tworzonych społecznościowo, zaś rozdział ósmy – klasyfikacji dokumentów tekstowych. W rozdziałach tych zacytowano źródła literaturowe z liczących się światowych konferencji naukowych i czasopism, w większości aktualne, choć brakuje tu zapewne źródeł najnowszych. Przegląd ten sprawia niestety wrażenie dokonanego w sposób dość wybiórczy i niesystematyczny, w szczególności brak tu szerszego omówienia aktualnych trendów badań – niemniej jednak można go uznać za wystarczający.

7. Jakie są wady i słabe strony rozprawy

Za główną słabość rozprawy uznać należy sposób ewaluacji zaproponowanych algorytmów. Dotyczy to zarówno doboru innych algorytmów i metod (lub ich braku), z którymi porównywane są te, które zostały zaproponowane przez Autora (dla przykładu – jednym z korpusów wykorzystanych w rozprawie jest zbiór pochodzący z konkursu BioASQ, rozprawa zawiera jednak tylko wynik przetwarzania zaledwie kilku dokumentów wybranych z tego zbioru), jak też i miar jakości działania algorytmów. O ile zastosowanie wyłącznie miar opartych na podobieństwie semantycznym (rozdział 7.2) i testowania hipotez statystycznych wydaje się być uzasadnione w przypadku algorytmu semantycznej kategoryzacji, to już przy porównywaniu metod opracowanych przez Autora z klasycznymi algorytmami klasyfikacji (rozdział 9) brak powszechnie stosowanych do ich ewaluacji miar (takich jak precyzja, zupełność, czy też krzywa ROC) jest znaczący. Prezentacja wyników eksperymentów (zastosowano wyłącznie formę tabelaryczną) także pozostawia nieco do życzenia. Wreszcie, rozprawa nie zawiera właściwie żadnej analizy dotyczącej złożoności obliczeniowej proponowanych algorytmów.

Zastrzeżenia budzi także opis innych algorytmów klasyfikacji semantycznej dokumentów tekstowych, czy też szerzej przegląd innych metod przetwarzania dokumentów i języka naturalnego, który wydaje się dość nieuporządkowany. Zwraca przy tym uwagę także sposób cytowania źródeł literaturowych – ich numeracja nie jest ani w porządku alfabetycznym, ani chronologicznym i wydaje

się być losowa.

Wreszcie rozprawa zawiera bardzo wiele ułomności, czy też wręcz błędów językowych (błędy gramatyczne, anglicyzmy i in.), które w niektórych przypadkach utrudniają nawet zrozumienie sensu jej treści.

8. Do której z następujących kategorii Recenzent zalicza rozprawę:

~~**a/ nie spełniająca wymagań stawianych rozprawom doktorskim przez obowiązujące przepisy**~~
~~**b/ wymagająca wprowadzenia poprawek i ponownego recenzowania**~~
c/ zadowalająco spełniająca wymagania
~~**d/ wyraźnie wykraczająca poza poziom przeciętny *spełniająca wymagania z nadmiarem)**~~
~~**e/ wybitna**~~

Reasumując stwierdzam, iż recenzowana rozprawa „Metody semantycznej kategoryzacji w zadaniach analizy dokumentów tekstowych" spełnia wymagania stawiane przez odnośną ustawę i tym samym wnoszę do Rady Naukowej Instytutu Podstaw Informatyki Polskiej Akademii Nauk o dopuszczenie mgr. Piotra Borkowskiego do dalszych etapów postępowania w przewodzie doktorskim.

Podpis