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

Warszawa, 4/6/2021

Recenzja rozprawy doktorskiej

Tytuł:
Wyznaczanie wielowymiarowego podobieństwa semantycznego w ontologiach
(Computing multidimensional semantic similarity in ontologies)

Autor rozprawy: **mgr inż Paweł Maciej Szmeja**

Promotor: **dr hab. Maria Ganzha prof. PAN**

Promotor pomocniczy: **dr Wiesław Pawłowski**

Rozprawa doktorska została wykonana w:

Instytut Badań Systemowych, Polskiej Akademii Nauk

Obliczenie podobieństwa jest bardzo ważnym zagadnieniem w technikach analizy danych takich jak klasyfikacji, grupowania danych lub w wyszukiwaniem informacji. Wraz z rozwojem narzędzi informatycznych do reprezentacji wiedzy takich jak Word Net, RDF, SPARQL, OWL oraz z powstaniem nowych idei takich jak Semantic Web, Semantic Search, Linked Open Data, Semantic Data Lakes, itp., problem mierzenia podobieństwa staje się coraz bardziej krytyczny. Rosnąca liczba semantycznie adnotowanych zasobów wymaga nowych metod i algorytmów przetwarzania, łączenia, porównywania różnych heterogenicznych zasobów. Recenzowana praca przedstawia przegląd istniejących metod mierzenia podobieństwa, w szczególności podobieństwa semantycznego, między złożonymi obiektami. Autor rozprawy, mgr Paweł Szmeja, zaproponował nową metodę

opisu podobieństwa, zwaną *wymiarowe podejście do obliczenia podobieństwa*, pozwalająca na pogrupowania istniejących metod podobieństwa. Zaprojektował i implementował również platformę SimDim zawierający ogólny algorytm obliczenia podobieństwa, zwłaszcza podobieństwa w ontologiach.

Przedstawiona do recenzji praca liczy 128 stron, nie licząc stron zawierających streszczenia oraz spis treści. Napisana została w języku angielskim. Praca została napisana bardzo starannie i przejrzyście, a jej układ nie budzi większych zastrzeżeń.

Zakres rozprawy

Pierwszym celem badawczym jest opracowanie metody opisu istniejących miar podobieństwa semantycznego w celu ich katalogowania i grupowania. Pomysł polega na podziale wiedzy na wymiary (aspekty) w kontekście podobieństwa semantycznego.

Następnym celem badawczym rozprawy jest definiowania modelu podobieństwa zawierającego ogólny algorytm wyznaczania podobieństwa semantycznego, stosowalnego niezależnie od domeny, struktury danych lub sposobu reprezentacji wiedzy, i zdolnego do uogólnienia istniejących metod. Kolejnym celem badawczym jest implementacja ogólnego algorytmu wyznaczania podobieństwa semantycznego z przykładami dla języka OWL.

W mojej ocenie, podjęta problematyka rozprawy jest bardzo ważna i aktualna. Cel pracy został też bardzo jasno i poprawnie sformułowany

Zawartość Rozprawy

Rozprawa doktorska Pana mgr. Pawła Szmeji składa się z 6 rozdziałów, jednego dodatku i spisu literatury.

Pierwszy rozdział zawiera motywacje i znaczenia rezultatów rozprawy w analizie danych. Autor opisuje cele badawcze i główne wkład rozprawy. W tym rozdziale, autor również podał podstawowe informacje na temat logiki deskrypcyjnej i język OWL.

Drugi rozdział przedstawia szeroki przegląd miar podobieństwa semantycznych, w szczególności podobieństwo w grafie, podobieństwo w taksonomii (ontologii), podobieństwo w logice lub podobieństwo w specyficznych dziedzinach takich, jak w Gene Ontology. W tym rozdziale autor przedstawił również dotychczasowe próby klasyfikacji i katalogowanie miar podobieństwa oraz istniejące systemy i oprogramowania wspomagające obliczenia podobieństwa semantycznego (HESML, SML, WNetSS, WordNet::Similarity, WS4J, NLTK, GO, GraphSim...).

W trzecim rozdziale autor wprowadził ideę wymiarów podobieństwa, które mogą być interpretowane jako różne aspekty podobieństwa dla złożonych obiektów. Autor wprowadził 3 główne wymiary: wymiar taksonomiczny, wymiar opisowy i wymiar leksykalny. Autor opisał również wymiar przynależnościowy (membership dimension) oraz wymiar współwystępowania (co-occurence dimension). Każdy wymiar może mieć również podwymiary, np. wymiar opisowy zawiera takie podwymiary jak wymiar fizyczny i wymiar kompozycyjny.

Rozdział czwarty zawiera szczegółowy opis generycznego algorytmu podobieństwa, który jest centralnym elementem platformy SimDim. Opracowany algorytm jest parametryzowany za pomocą funkcji cechujących, funkcji podobieństwa oraz reduktorów. Zawartość tego rozdziału jest jednym z głównych rezultatów rozprawy. Oprócz precyzyjnych opisów matematycznych, autor podał również ilustrujące przykłady dla metod przetwarzania języka naturalnego z użyciem ontologii Word Net.

Rozdział piąty przedstawia implementację w języku programowania Skala ogólnego algorytmu wyznaczania podobieństwa semantycznego wraz z przykładami dla języka OWL.

Rozdział szósty podsumuje zawartość rozprawy i przedstawia otwarte problemy i możliwe kierunki dalszych badań.

Rozprawa doktorska zawiera również dodatek, w którym autor zapisuje spis 28 funkcji cechujących, 31 funkcji podobieństwa i 23 reduktorów stosowanych w platformie SimDim.

Istotne elementy rozprawy

W mojej ocenie, przedstawiona do recenzji praca jest najbardziej kompletnym kompendium wiedzy na temat podobieństwa semantycznego. Pod tym względem, rozprawa jest istotnym rozszerzeniem istniejących w literaturze prac przeglądowych.

Rozprawa doktorska przedstawia również autorskie podejście do zagadnienia podobieństwa semantycznego. Pan mgr Paweł Szmeja zaproponował ideę wymiarów podobieństwa, które służą do grupowania i katalogowania istniejących miar podobieństwa semantycznego i stanowią podstawę dla platformy DimSim. Trzy główne wymiary: wymiar taksonomiczny, wymiar opisowy i wymiar leksykalny, które mogą być interpretowane jako niezależne (prostopadłe) aspekty, definiują przestrzeń możliwych złożonych miar podobieństwa.

Zaprojektował i implementował również bardzo bogatą platformę DimSim, która jest narzędziem do budowy systemu podobieństwa semantycznego dla specyficznej ontologii dziedzinowej. Przykład implementacji dla OWL został też bardzo dokładnie opisany i udostępniony na Github.

Praca została bardzo starannie przygotowana pod względem edytorskim. Struktura pracy jest logiczna, a sformułowanie tezy i opisy metod są bardzo klarowne. Autor wykazał dobrą technikę pisania prac naukowych. Praca została napisana w dobrym, technicznym języku angielskim i nie zauważyłem istotnych błędów ortograficznych.

Pan mgr Paweł Szmeja jest współautorem dwóch publikacji związanych z tematyką rozprawy.

Uwagi krytyczne

Moja pierwsza uwaga dotyczy kluczowego pojęcia dla rozprawy, tj. miara podobieństwa. Brakuje na samym początku rozprawy określenia własności dla miar podobieństwa, np. czy miara podobieństwa może być ujemna, czy też nieskończona, kiedy podobieństwo między dwoma obiektami jest większe niż dla innej pary, istotnie utrudnia zrozumienie uwag i dyskusji w pierwszych trzech rozdziałach. Pierwsza matematyczna definicja pojęcia funkcji podobieństwa pojawi się dopiero na stronie 61, ale zostało używane dość często wcześniej.

Mimo, że rozprawa doktorska jest z zakresu informatyki technicznej, uważam, że autor powinien przeprowadzić dyskusje na temat własności proponowanego rozwiązania. W przypadku algorytmów, warto przeprowadzić analizy złożoności obliczeniowej. Np. jaka jest złożoność algorytmu obliczenia podobieństwa w grafach, w taksonomiach lub w logice. Jeśli algorytmy te zostały implementowane w systemie, jaka była złożoność empiryczna, itp.

Bardzo ciekawie wyglądają również pojęcia „najodpowiedniejsza funkcja podobieństwa" i „najodpowiedniejsze cechowanie". Szkoda, że autor podał jedynie definicje tych pojęć bez zbadania ich własności.

Oceniam, że dorobek publikacji autora rozprawy jest zbyt skromny. Pan mgr Paweł Szmeja jest współautorem dwóch artykułów o samym tytule. Prawdopodobnie, drugi artykuł jest rozdziałem w książce będącym rozszerzeniem pierwszego artykułu.

Wnioski końcowe

Mgr Paweł Szmeja wykazał się odpowiednią wiedzą z zakresu zarządzania i modelowania bazą wiedzy, umiejętnością programowania, oraz zdolnością do modelowania pomysłów. Spora część uwag krytycznych wynikała z faktu, że rozprawę doktorską przygotowano w formie raportu do projektu, i brakuje analizy proponowanego rozwiązania. W mojej ocenie, największą słabością rozprawy jest słaby dorobek publikacji.

Z drugiej strony, uważam, że merytoryczna zawartość rozprawy stanowi ciekawy i oryginalny wkład w rozwój praktycznych metod semantycznej analizy i przetwarzania danych. Jest to 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 rozprawa doktorska p.t. „Wyznaczanie wielowymiarowego podobieństwa semantycznego w ontologiach" autorstwa Pawła Szmeji spełnia wymogi stawiane pracom doktorskim. Wnoszę zatem o dopuszczenie przedłożonej mi do recenzji rozprawy do publicznej obrony

Nguyen Hung Son