Streszczenie rozprawy doktorskiej

Wyznaczanie wielowymiarowego podobieństwa semantycznego w ontologiach

(Computing multidimensional semantic similarity in ontologies)

mgr inż. Paweł Maciej Szmeja

Opieka naukowa:
dr hab. Maria Ganzha prof. PAN
dr Wiesław Pawłowski

Motywacja

Istnieją różne sposoby na mierzenie podobieństwa, z których wiele sięga poza matematykę i informatykę, zagłębiając się w psychologię, socjologię, a nawet filozofię. W przeszłości, ze względu na swoje filozoficzne korzenie, pojęcie podobieństwa semantycznego odnosiło się do bliskości *znaczenia słów*. W miarę upływu lat to pojęcie zostało rozszerzone o podobieństwo zdań, dokumentów, a nawet całych zbiorów dokumentów. W tym obszarze, jako podstawa obliczeń i źródło wiedzy na temat języka naturalnego, używane były słowniki i tezaurusy. Ontologie, będące naturalnym rozszerzeniem idei słowników, były również coraz częściej wykorzystywane, w miarę wzrostu ich ogólnej popularności w informatyce. Dziś, ontologie i tezaurusy, takie jak WordNet są często używane do obliczania podobieństwa słów, zdań i dokumentów. Inne, specyficzne dla konkretnych obszarów zainteresowań (domen) ontologie wytyczają przestrzenie, w której obliczanie podobieństwa semantycznego jest niezależne od znaczenia słów, a raczej operuje na wysoce wyspecjalizowanej wiedzy przechowywanej w ontologii. Rozpowszechnienie ontologii rozszerzyło znaczenie podobieństwa semantycznego (poczynając od *znaczenia słów*), o znaczenie *bytów i obiektów ontologicznych*.

Nowy, praktyczny, wymiar został nadany semantyce wraz z rozwojem narzędzi informatycznych takich, jak RDF, SPARQL, OWL, triplestores, a także podejść i idei w rodzaju Semantic Web, Linked Open Data, semantic data lakes, itd. Wraz z powstaniem technologii, które w praktyczny sposób umożliwiają semantyczną reprezentację i przetwarzanie informacji, problem podobieństwa zaczął być rozpatrywany w nowym, informatycznym, kontekście. Przechodząc od teorii do praktyki, można zaobserwować, że wraz z rosnącą liczbą semantycznie opisanych zasobów, rośnie potrzeba opracowania nowych metod i algorytmów do ich przetwarzania. Jedną z najważniejszych właściwości danych semantycznych jest możliwość łączenia wielu heterogenicznych zasobów. Obliczanie podobieństwa jest co najmniej użyteczne, a czasami nawet niezbędne w procesach wyszukiwania wzajemnych powiązań, wnioskowaniu na podstawie istniejących informacji, lub, pisząc bardziej ogólnie, zautomatyzowanej interakcji z danymi semantycznymi.

Obliczanie podobieństwa bytów w ontologiach umożliwia wyszukiwanie, filtrowanie i inteligentne przetwarzanie semantycznie adnotowanych danych. Jest też fundamentem dla wielu inteligentnych aplikacji, algorytmów i rozwiązań dla różnorakich form zarządzania danymi, które coraz częściej korzystają z semantyki.

Pomimo znaczącego rozwoju technologii semantycznych, ogólne metody obliczania podobieństwa (tzw. miary podobieństwa) pozostają nieco w tyle. Analiza stanu wiedzy ujawnia szereg

problemów z obecnie dostępnymi metodami obliczania podobieństwa w ramach ontologii. Metody, które bardzo głęboko analizują wiedzę ontologiczną są wysoce wyspecjalizowane i dostępne wyłącznie dla bardzo specyficznych ontologii domenowych. Ponieważ zakładają one określoną strukturę ontologii lub istnienie określonych właściwości i adnotacji, nie mogą być używane w dowolnych ontologiach. Z drugiej strony, metody, które są stosowalne dla dowolnych ontologii, często zdefiniowane na wysokim poziomie abstrakcji, nigdy nie wykorzystują pełnego zakresu dostępnej wiedzy (lub ekspresywności logicznej), bardzo często ograniczając się do taksonomii. Te i inne ograniczenia sugerują możliwość poprawy stanu wiedzy, niekoniecznie jeśli chodzi o wydajność implementacji oprogramowania, ale raczej lepsze (pełniejsze) wykorzystanie wiedzy zawartej w ontologiach.

Zakres i cel pracy

W obszarach, takich jak Internet Rzeczy (IoT), dane, które często są najbardziej interesujące dla badaczy i najważniejsze dla użytkowników, to dane „transakcyjne" (na przykład dane obserwacji lub akcji, które tracą znaczenie w czasie), lub dotyczą instancji (na przykład informacje o konkretnych sensorach). Tradycyjne miary podobieństwa, ze względu na swoje skupienie na taksonomii, mają ograniczoną użyteczność w wielu zastosowaniach IoT, ponieważ charakteryzują się niewrażliwością na zmiany w danych transakcyjnych.

Kolejnym problemem jest domniemana uniwersalność wyniku obliczania podobieństwa. Różne metody przedstawiają różne, nawet w skrajnym stopniu, wyniki. Stosując różne metody można spotkać się z sytuacją, w której jedna metoda nie rozpoznaje podobieństwa w ogóle, a inne stwierdzają znaczące podobieństwo. Pomimo tego faktu, mówi się, że wszystkie miary obliczają „to samo" *semantyczne* podobieństwo, które funkcjonuje jako pojęcie uniwersalne, stosowalne do dowolnego wyniku dowolnej miary podobieństwa. Mimo różnego rozumienia i sposobu obliczania podobieństwa proponowanych przez różnych autorów, wydaje się, że celem jest zawsze odwzorowanie jakiegoś teoretycznego, idealnego podobieństwa. Wynik podobieństwa nie ma więc dodatkowej interpretacji i nie odzwierciedla różnicy w podejściach, a jedyną informacją (ewentualnie) dołączoną do wyniku jest nazwa użytej miary.

W świetle wyżej wymienionych zagadnień, związanych z interpretacją wyników podobieństwa i powstałych dotychczas współczesnych miar, potrzebne jest uaktualnione i zmodernizowane podejście do obliczania podobieństwa, w tym podobieństwa w ontologii. Wobec tego zdefiniowano następujące cele badawcze:

1. Stworzenie sposobu opisu podobieństwa niezależnego od dotychczasowych podejść do klasyfikacji i grupowania modeli i miar podobieństwa. Podobieństwo powinno mieć, jasną interpretację i znaczenie obejmujące dotychczas stosowane miary. Podejście powinno być stosowalne do wyników metod obecnie istniejących, jak i tych utworzonych w przyszłości.

2. Definicja modelu podobieństwa, zawierającego ogólny algorytm wyznaczania podobieństwa semantycznego, stosowalnego niezależnie od domeny, struktury danych, lub sposobu reprezentacji wiedzy, i zdolnego do generalizacji istniejących metod.

3. Definicja i implementacja ogólnego algorytmu wyznaczania podobieństwa semantycznego, razem z przykładami dla języka OWL.

Podsumowanie zawartości rozprawy

Jako narzędzie do grupowania istniejących metod podobieństwa, w pracy zaproponowane zostało *wymiarowe podejście do obliczania podobieństwa*, wypełniające warunki stawiane w pierwszym celu badawczym. Przedstawia ono teoretyczny model podziału wiedzy na wymiary, w kontekście podobieństwa semantycznego. Wprowadzenie pojęcia wymiaru podobieństwa oferuje nowy kontekst w którym analizowane i porównywane mogą być różne metody. Dodatkowa informacja na temat „wymiarowości" podobieństwa nadaje wynikom wyraźną interpretację, według której wyniki dostarczają informacji na temat różnych aspektów podobieństwa. Pozwala to na wyraźne rozróżnienie i wytłumaczenie rozbieżnych wyników różnych metod, oraz bardziej świadomy wybór metody obliczania podobieństwa dla dowolnego zadanego problemu.

Rozprawa przedstawia również wyniki badań nad podobieństwem semantycznym, ze szczególnym wskazaniem na podobieństwo w ontologiach. Oprócz analizy stanu wiedzy, głównym wkładem jest platforma („framework") podobieństwa *SimDim* zawierająca ogólny algorytm podobieństwa, będący realizacją nowego cechowego modelu podobieństwa (realizując tym samym drugi cel badawczy), oraz opis *wymiarowego* podejścia do podobieństwa, które jest realizowalne za pomocą wspomnianego algorytmu. Platforma *SimDim* jest w stanie włączyć w obliczanie podobieństwa wiedzę często pomijaną w innych podejściach do obliczania podobieństwa, a tym samym wzbogacenie aktualnego stanu wiedzy. *SimDim* oferuje wysoki stopień konfiguracji, a dobór parametrów konfiguracyjnych ma duży wpływ jej właściwości, zachowanie i wydajność.

Generyczny algorytm podobieństwa semantycznego wraz z implementacją i przykładami dla języka OWL, wchodzący w skład platformy *SimDim*, jest oparty na podejściu cechowym i realizuje trzeci cel badawczy. Ze względu na definicję na wysokim poziomie abstrakcji jest on stosowalny do dowolnej domeny, a nawet do danych spoza ontologii (choć rozprawa skupia się na podobieństwie bytów ontologicznych). Dobranie różnych parametrów algorytmu zmienia jego zachowanie w znaczący sposób, a różne parametry tworzą osobne *instancje SimDim*, pozwalając na realizację w ramach *SimDim* innych istniejących podejść. Rozprawa zawiera również opis wielu przykładowych parametrów (będących funkcjami) i instancji.