prof. dr hab. inż. Krzysztof Goczyła, prof. zw. PG
Politechnika Gdańska
Wydział Elektroniki, Telekomunikacji i Informatyki
kris@eti.pg.edu.pl

Gdańsk, 31.01.2018

Recenzja rozprawy doktorskiej

mgr. inż. Jędrzeja Potońca

zatytułowanej:

Methods for Automatic Enrichment of Ontologies from Linked Data
(Metody automatycznego rozszerzania ontologii na podstawie powiązanych danych)

1. Problem badawczy i jego znaczenie

Tematem rozprawy jest analiza możliwości tworzenia nowych lub wzbogacania istniejących ontologii z otwartych źródeł internetowych zapisanych w formacie zwanym powiązanymi danymi (ang. *Linked Data*). Format ten opiera się na języku RDF – bardzo ogólnym formalizmie stworzonym do zapisu informacji z różnych dziedzin techniki i życia codziennego. Aktualnie w Internecie można znaleźć sporo użytecznych zasobów tego typu, ogólnych - jak na przykład DBpedia (semantyczny odpowiednik Wikipedii), a także specjalnych, jak na przykład biblioteki cyfrowe, a także zasobów tworzonych przez szeregowych użytkowników Internetu.

Powiązane dane – w swojej naturalnej postaci - są jednak trudne w bezpośredniej interpretacji nie tylko przez człowieka, ale także przez komputery z uwagi na mnogość występujących w nich pojęć i złożoność wzajemnych powiązań pomiędzy tymi pojęciami. Repozytoria powiązanych danych są czasem zaopatrzone w ogólny opis – ontologię wysokiego poziomu, która jednak daje tylko ogólne pojęcie o danym zasobie. Istotnym problemem jest więc wzbogacanie tego opisu w taki sposób, by wydobyć z powiązanych danych ukrytą w nich, czasem bardzo głęboko, wiedzę na temat dziedziny, której dotyczą. Umożliwia to zarówno ludziom, jak i komputerom lepsze wykorzystanie tych zasobów, na przykład w inżynierii systemów informatycznych.

Autor postawił przed sobą niełatwy problem częściowego zautomatyzowania procesu wydobywania ukrytej wiedzy z powiązanych danych. Bez wątpienia jest to problem o charakterze naukowym, wpisujący się w aktualną tematykę badawczą z obszaru zarządzania wiedzą i z obszaru Semantycznego Internetu (ang. *Semantic Web*). W zamierzeniu autora taki proces ma w sposób istotny wspomóc inżyniera ontologii w tworzeniu i rozwoju ontologii użytecznych w inżynierii systemów opartych na wiedzy, przeznaczonych do konkretnych zastosowań praktycznych. W tym kontekście tematyka rozprawy ma potencjalnie duże znaczenie praktyczne, szczególnie zważywszy na fakt, że w aktualnym stanie Semantycznego Internetu tempo tworzenia zasobów danych znacznie wyprzedza tempo ich inteligentnego wykorzystywania. Należy się spodziewać, że ten proces może się pogłębiać w związku ze zjawiskiem zwanym *Big Data* i urzeczywistnianiem idei Internetu Rzeczy, choć w tej chwili trudno jest przewidzieć, w jakich formatach i w jakim zakresie udostępniane będą dane tego typu.

2. Wkład autora

Za główny wkład rozprawy w obszar badawczy opisanym w punkcie 1. niniejszej recenzji jest opracowanie i weryfikacja trzech metod półautomatycznego i automatycznego rozszerzania ontologii o ekspresywności języka OWL 2 EL z powiązanych danych. Są to oryginalne metody autorskie, które zostały zaimplementowane albo całkiem od nowa, albo z wykorzystaniem istniejących środowisk. Metody te zostały opisane w rozdziałach 4., 5. i 6. rozprawy. Pierwsza metoda, opisana w rozdziale 4., wykorzystuje programowanie całkowitoliczbowe do wzbogacania istniejącej hierarchii klas o nowe koncepty bazowe i o nowe koncepty złożone, definiowane za pomocą konceptów bazowych i innych konceptów złożonych. Metoda ta nadaje się do analizy powiązanych danych opisanych bardzo ubogimi ontologiami, zawierającymi jedynie koncepty bardzo ogólne („oczywiste") i bardzo proste zależności (role) pomiędzy konceptami. Metodę tę zaimplementowano w RMonto, wtyczce do środowiska RapidMiner służącego do eksploracji danych.

Druga metoda, opisana w rozdziale 5., odnosi się do bardzo częstych przypadków, jakimi są powiązane dane opisane bardzo płytkimi, płaskimi ontologiami. Pozwala ona na wykrywanie ukrytych zależności podrzędności (subsumpcji) pomiędzy konceptami, a oparta jest na teorii matematycznej zwanej formalną analizą pojęć. Metoda opracowana przez autora w zamierzeniu wymaga interakcji z ekspertem, co należy raczej uznać za jej zaletę niż wadę z uwagi na to, że zastosowana formalna analiza pojęć może generować wiele aksjomatów nieużytecznych lub bardzo skomplikowanych, które zamiast wzbogacić wiedzę o danym zasobie, mogą tę wiedzę zakryć trudnymi do zrozumienia wyrażeniami konceptowymi. Metoda została zaimplementowana w języku Java.

Trzecia metoda, opisana w rozdziale 6., nazwana *Swift Linked Data Miner* (SLDM), wykorzystuje algorytm odkrywania częstych wzorców w zasobie powiązanych danych, stosując w tym celu zapytania w języku SPARQL odpowiadające poszczególnym wyrażeniom konceptowym języka OWL 2 EL. Autor zastosował tu oryginalne podejście polegające na systematycznym próbkowaniu zbioru powiązanych danych i wykrywaniu częstych wzorców w strukturze danych odpowiednio zaprojektowanej do przechowywania i analizowania trójek RDF. Metoda ta również została zaimplementowana w postaci biblioteki i wtyczek do popularnych środowisk edycji ontologii z rodziny Protégé.

Wymieniony wyżej oryginalny wkład autora uważam za oryginalny i wartościowy dla informatyki w obszarze zarządzania wiedzą. Jednak podczas studiowania rozprawy nasunęło mi się szereg zastrzeżeń i uwag o charakterze dyskusyjnym dotyczących opracowanych i zaprezentowanych w rozprawie metod.

1. W rozdziale 4. brakuje jakiegokolwiek opisu eksperymentów z rzeczywistymi zbiorami powiązanych danych. Przedstawiono jedynie narzędzie i proste przykłady ilustrujące metodę. Niestety nie zostały zaprezentowane jakiekolwiek nowe koncepty bazowe ani nowe wyrażenia konceptowe wydobyte z rzeczywistych źródeł powiązanych danych. Ta sama uwaga dotyczy metody opisanej w rozdziale 5. Szczegółowo zaprezentowano narzędzie, załączając nawet zrzuty ekranów, jednak zabrakło przytoczenia wyników eksperymentów.

2. W metodach opisanych w rozdziale 4. i 5. autor ogranicza się do atrybutów binarnych. Jest to poważne ograniczenie, gdyż w praktyce o wiele częściej występują atrybuty wyliczeniowe (np. atrybut „kolor", a nie atrybut „kolor czerwony") i atrybuty ilościowe. Czy opisane metody można uogólnić na atrybuty inne niż binarne?

3. W metodzie SLDM, a także – domyślnie – w pozostałych dwóch metodach autor stosuje założenie o świecie zamkniętym (*Closed World Assumption*, CWA) (patrz. np. formuły 6.25 i 6.26), podczas gdy wiadomo, że ontologie opisane językiem OWL i silniki wnioskujące przyjmują założenie o świecie otwartym (*Open World Assumption*, OWA). Przydałaby się jakaś dyskusja tej kwestii.

4. W rozdziale 6 znajduje się podrozdział 6.7 „Experimental evaluation", po którym obiecywałem sobie prezentacji przykładów wykrytych wzorców z istniejących źródeł powiązanych danych. Jednak zawiodłem się, gdyż w tym podrozdziale zawarto opis eksperymentu crowdsourcingowego, koncentrującego się na analizie skuteczności i poprawności w wykrywaniu konceptów w znanej ontologii, a nie na pokazaniu wykrytych nowych konceptów i wyrażeń konceptowych. Rozumiem, ze przeprowadzenie pełnego, nietrywialnego eksperymentu byłoby trudnym wyzwaniem, jednak oczekiwałem choć podania aksjomatów zaproponowanych przez algorytm SLDM.

5. Autor nie przeprowadził nawet przybliżonej analizy złożoności obliczeniowej zaproponowanych metod. W rozdziale 6. dokonuje jakościowej analizy prowadzącej do wybrania odpowiedniej metody próbkowania i opisuje przeprowadzone eksperymenty ilościowe dotyczące czasu obliczeń i zajętości pamięci w konkretnym systemie komputerowym. Przydałaby się jednak jakaś analiza teoretyczna, choćby bardzo przybliżona.

6. Czy autor rozważał rozszerzenie ekspresywności rozszerzanych ontologii do innych dialektów języka OWL 2, szczególnie do OWL 2 DL? Jasne jest, że wybór dialektu EL powodowany był kwestiami złożonościowymi, jednak ma to znaczenie dopiero przy dużych ontologiach. W praktyce nawet niewielkie ontologie, rzędu kilkudziesięciu konceptów i ról, bardzo wzbogaciłyby naszą wiedzę o semantyce zasobów powiązanych danych, jeśli dopuścilibyśmy na przykład negację i sumę klas. Przykładowo, bardzo ważnym elementem wiedzy o świecie są aksjomaty pokrycia jednego konceptu przez koncepty rozłączne, których to aksjomatów nie można sformułować w dialekcie EL.

7. Z uwagi na brak wyników eksperymentalnych dla rzeczywistych zbiorów powiązanych danych trudno jest wskazać obszary zastosowań praktycznych proponowanych metod. Dla jakich klas czy też typów zbiorów powiązanych danych najlepiej działają metody? Czy autor zastanawiał się nad możliwością pojawienia się w wyniku wzbogacania ontologii definicji cyklicznych[1], szczególnie w metodzie opisanej w rozdziale 5., i nad konsekwencjami takich definicji?

Powyższe uwagi, poza uwagami dotyczącymi braku dostatecznej eksperymentalnej walidacji opracowanych metod, mają w dużym stopniu charakter dyskusyjny, dlatego interesujące będzie odniesienie się do nich przez autora rozprawy w trakcie publicznej obrony.

3. Poprawność

Rozprawa napisana jest w dobrym języku angielskim. W całym tekście natrafiłem na jedynie kilka błędów, głównie literowych. Znamienne, że załączone streszczenie w języku polskim nie jest już tak starannie napisane pod względem językowym i redakcyjnym i niestety zawiera sporo błędów typowych dla prac technicznych pisanych po polsku. Szczególnie drażni maniera nadużywania wielkich liter nie tylko w nazwach pojęć pospolitych, jak np. formalna analiza pojęć, ale także

[1] Zob. K. Goczyła, *Ontologie w systemach informatycznych*, EXIT 2011, rozdział 6.5.

w odnośnikach (np. „na Rysunku 3” zamiast „na rysunku 3”) i w pisowni wielkości fizycznych (np. „10GB” zamiast „10 GB”). Są to wszystko anglicyzmy, które nie powinny mieć miejsca w poważnych pracach naukowych pisanych w języku polskim. Jednak moją recenzję w jej aspektach merytorycznych w całości opieram na wersji anglojęzycznej.

Warto podkreślić, że praca jest bardzo starannie zredagowana pod względem notacji matematycznej i formy prezentacji wyników przeprowadzonych analiz jakościowych i ilościowych. W paru przypadkach miała jednak miejsce nadmierna formalizacja. I tak na przykład na stronie 48 autor wyprowadza kilka formuł matematycznych (od 6.1 do 6.4) dla opisu prostej i oczywistej operacji, jaką jest podział dużego zbioru danych na kilka rozłącznych części o zadanej maksymalnej wielkości. Zdarza się też stosowanie tego samego symbolu do oznaczenia różnych wielkości (np. symbol *M* wprowadzony w rozdziale 4, oznaczający według wykazu pojęć pewną dużą liczbę, w podrozdziale 5.5 oznacza zbiór wyrażeń konceptowych).

4. Wiedza kandydata

Moim zdaniem, autor rozprawy wykazał się bardzo dobrą wiedzą w zakresie Semantycznego Internetu i inżynierii ontologii oraz języka SPARQL stanowiącego standard dostępu do powiązanych danych. Przedstawił również w sposób wyczerpujący aktualny stan wiedzy w zakresie wydobywania wiedzy ontologicznej z zasobów Semantycznego Internetu, wskazując na niedostatki istniejących metod i bardzo ograniczony zakres ich stosowalności. Swoją znajomość obszaru objętego tematyką rozprawy potwierdził szerokim doborem literatury przedmiotu. Ważne jest także to, że p. Jędrzej Potoniec opublikował wyniki swoich prac w renomowanych czasopismach i na konferencjach międzynarodowych z obszaru Semantycznego Internetu i sztucznej inteligencji (choć układ spisu literatury nie ułatwił mi wydobycia tych informacji).

Generalnie stwierdzam, że kandydat posiadł ogólną wiedzę w dyscyplinie informatyka wystarczającą do prowadzenia samodzielnych badań w tej dyscyplinie w obszarach badawczych związanych z zarządzaniem wiedzą.

5. Podsumowanie

Pan mgr inż. Jędrzej Potoniec osiągnął oryginalne i wartościowe wyniki naukowe w obszarze zarządzania wiedzą mieszczącym się w zakresie nowoczesnych badań informatycznych. Wnoszę więc o przekazanie recenzowanej rozprawy do dalszych etapów przewodu doktorskiego.

Biorąc pod uwagę opinie zaprezentowane w poprzednich punktach i wymagania zdefiniowane przez art. 13 Ustawy z dn. 14 marca 2003 r. o stopniach naukowych i tytule naukowym (z późn. zmianami)[2] moja ocena rozprawy pod względem trzech podstawowych kryteriów jest następująca:

A. Czy rozprawa zawiera oryginalne rozwiązanie problem naukowego? (wybierz jedną opcję stawiając znak **X)**

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

[2] http://www.nauka.gov.pl/g2/oryginal/2013_05/b26ba540a5785d48bee41aec63403b2c.pdf

B. Czy po przeczytaniu rozprawy zgadzasz się, że kandydat posiada ogólną wiedzę teoretyczną w dyscyplinie informatyka?

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

C. Czy kandydat posiadł umiejętność samodzielnego prowadzenia pracy naukowej?

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

podpis