Prof. dr hab. Henryk Rybinski
Instytut Informatyki PW
Warszawa, Nowowiejska 17

Warszawa, 2018-03-30

Recenzja
rozprawy doktorskiej mgra inż Jędrzeja Potońca
pt. "Methods for Automatic Enrichment of Ontologies
from Linked Data".

Wstęp

Przedstawiona do recenzji rozprawa jest napisana w języku angielskim. Składa się z siedmiu rozdziałów, bibliografii, trzech dodatków oraz indeksu. Objętość rozprawy - 105 stron w tym 83 strony tekstu podstawowego. W swojej opinii przedstawię ogólną charakterystykę rozprawy, a następnie przedstawię swoje uwagi ogólne i szczegółowe.

1. Problem Badawczy

Obszarem badań zaprezentowanych w rozprawie jest dziedzina odkrywania wiedzy w zasobach informacyjnych. Zagadnienie to wiąże się dość mocno z kilkom innymi współczesnymi problemami informatyki, takimi jak metody reprezentacji wiedzy, metody eksploracji danych, czy wreszcie zagadnienia uczenia maszynowego. Precyzując obszar zainteresowań doktoranta – rozprawa dotyczy metod tworzenia i utrzymywania ontologii w oparciu o zasoby informacyjne pajęczyny semantycznej (ang. *semantic web*). Badania związane z ontologiami w kontekście sztucznej inteligencji wiążą się ściśle z metodami reprezentacji wiedzy i towarzyszą badaniom w dziedzinie sztucznej inteligencji od pierwszej konferencji poświęconej AI. Pierwszy raz termin „ontologia" w tym kontekście pojawia się w roku 1980, kiedy to McCarthy stwierdził, że budując systemy inteligentne bazujące na logice trzeba "*list everything that exists, building an ontology of our world*".

W kontekście systemów informacyjnych tematyka ontologii pojawia się na początku lat 90-tych, natomiast w kontekście semantycznej pajęczyny w roku 2002, wraz z propozycją autorstwa Tim Berners-Lee. Podstawowym problemem szerokiego stosowania ontologii w praktyce jest jednak złożoność ich budowania i utrzymywania.

Dlatego od ok. 20 lat trwają intensywne badania nad automatyzacją tworzenia ontologicznych baz wiedzy z różnych zasobów informacyjnych. Wykorzystywane są w tym celu metody eksploracji danych, uczenia maszynowego, czy też ogólniej - odkrywania wiedzy. Wykorzystywane do tego celu zasoby informacyjne, to przede wszystkim zasoby internetowe, takie jak elektroniczne korpusy tekstów (podręczniki, encyklopedie), słowniki i tezaurusy, schematy baz danych, a od ok. 10 lat także zasoby internetowe związane z ideą pajęczyny semantycznej (*semantic web*).

Opiniowana rozprawa leży w nurcie badań związanych z odkrywaniem wiedzy z internetowych zasobów danych powiązanych tworzących pajęczynę (*linked (open) data*). W szczególności przedmiotem rozprawy są metody wspomagania inżyniera wiedzy przy tworzeniu i utrzymywaniu ontologii.

W rozprawie autor opracowuje trzy nowe metody. Pierwsza dotyczy odkrywania nowych pojęć w danych powiązanych (*linked data*). Metoda druga bazuje na Formalnej Analizie Pojęć i pozwala generować aksjomaty tworzące taksonomię pojęć. Metoda trzecia jest oparta na algorytmach odkrywania wzorców.

Podsumowując, uważam, że tematyka rozważana przez doktoranta z praktycznego, jak też teoretycznego punktu widzenia jest ważna i jest godna rozprawy doktorskiej.

Konstrukcja rozprawy

W konstrukcji pracy wyróżnić można następujące zasadnicze części:

1. motywacja badań, omówienie problemów badawczych poruszanych w pracy (rozdziały 1, 2),
2. przegląd literatury (Rozdział 3),
3. prezentacja opracowanych przez doktoranta nowych metod automatycznego wspomagania budowania i utrzymywania ontologii (rozdziały 4 - 6);
4. podsumowanie i wnioski (Rozdział 7);
5. Załącznik z pełnym opisem OWL 2 EL.

Charakter rozprawy

Zawartość rozprawy dowodzi solidnego teoretycznego warsztatu doktoranta w zakresie metod reprezentacji wiedzy, tworzenia i utrzymywania ontologii, metod eksploracji danych czy też uczenia maszynowego, jednocześnie badania autora są wsparte implementacjami, te zaś wskazują na

1. równie solidny warsztat doktoranta w zakresie metod i narzędzi związanych z wyszczególnionymi wyżej dziedzinami
2. możliwości praktycznego zastosowania opracowanych algorytmów.

2. Wkład autora

Wkład autora oceniam wysoko – obejmuje on szereg ważnych elementów związanych z odkrywaniem wiedzy na potrzeby budowania i utrzymywania ontologii. Wyróżnić tu można dwie metody opracowane przez autora:

1. Metoda odkrywania nowych pojęć pojawiających się w danych powiązanych (Rozdz. 4);
2. Metoda odkrywania wzorców, pozwalająca wykrywać wyrażenia klasy OWL 2 EL z danych powiązanych (Rozdz. 6);

oraz jedną metodę zmodyfikowaną przez doktoranta:

3. Metoda generowania aksjomatów tworzących taksonomię pojęć (Rozdz. 5).

Ad. 1

Opracowana metoda polega na wygenerowaniu zbioru kandydatów a następnie sprowadzenie wyboru najbardziej interesującego podzbioru kandydatów do problemu optymalizacji, który można rozwiązywać metodami programowania całkowito-liczbowego.

Ad. 2

Opracowana została przez doktoranta oryginalna metoda o nazwie Swift Linked Data Miner (SLDM). Jest to metoda bazująca na odkrywaniu wzorców. Jej zadaniem jest eksploracja wyrażeń klas w OWL2EL i odkrywanie aksjomatów dla wybranych klas. Zaproponowane są metody pobierania istotnych danych z odległych zasobów danych powiązanych oraz sposób ich reprezentowania w formie trzypoziomowego indeksu. Przedstawione są algorytmy przetwarzające ten indeks w celu odkrywania wzorców. W efekcie uzyskuje się częściowe definicje klasy wskazanej przez użytkownika. Koncepcja algorytmu SLDM jest wsparta eksperymentem wykazującym, że wygenerowane w ten sposób aksjomaty są poprawne w szerokim zakresie ustawień parametrów. Doktorant przedstawił także własności obliczeniowe algorytmu i zaimplementował algorytm w formie wtyczki do Protege.

Ad. 3

Na bazie znanej teorii Formalnej Analizy Pojęć (*Formal Concept Analysis*) doktorant wprowadza rozszerzenia metod opracowanych przez F. Baadera i innych (poz. [5]). Opracowany algorytm generuje propozycje aksjomatów zawierania klas, które nie stoja w sprzeczności do istniejącej wersji ontologii. W zaproponowanym rozszerzeniu zastosowano klasyfikator, który uczy się z decyzji użytkownika. Zaimplementowana została aplikacja w języku Java.

Wszystkie trzy opracowane algorytmy mogą stanowić istotne wsparcie dla inżyniera wiedzy. Najważniejszym osiągnięciem doktoranta jest moim zdaniem algorytm SLDM.

3. Poprawność

Wysoko oceniam przyjętą przez autora metodologię badań.

Problem badawczy został prawidłowo zaprezentowany w p. 1.2. Autor dokonuje krytycznej analizy stanu badań w dziedzinie metod odkrywania wiedzy na potrzeby wzbogacania ontologii, w oparciu o tę analizę proponuje szereg własnych rozwiązań, jednocześnie przeprowadza badania eksperymentalne. Autor stworzył szereg narzędzi, które mogą być wykorzystywane w praktyce inżyniera wiedzy.

4. Wiedza kandydata

Wiedzę kandydata oceniam wysoko. Jest to przede wszystkim wiedza w zakresie metod reprezentacji wiedzy, budowy i utrzymywania ontologii, danych powiązanych (semantyczna pajęczyna), a także dziedzin towarzyszących zagadnieniom baz wiedzy, takich jak logika, czy też formalna analiza pojęć (*Formal Concept Analysis*), Przegląd literatury obejmuje 103 pozycje (w tej liczbie znajduje się 9 prac współautorstwa doktoranta, z czego w 6-ciu pracach doktorant występuje jako pierwszy autor). Wprowadzenie w problematykę oraz przegląd literatury jest przede wszystkim zawarty w rozdziałach 2 i 3. Mam pewne uwagi co do przeglądu literatury (kolejny punkt), jednak nie wpływają one na zasadniczą (wysoką) ocenę wiedzy doktoranta.

5. Inne uwagi

Praca jest napisana starannie i dobrym językiem angielskim (kilka drobnych poprawek językowych zaznaczyłem w dokumencie PDF). Układ pracy jest logiczny i podporządkowany tezie rozprawy. Zamieszczenie w rozprawie listy stosowanych symboli oraz indeksu ułatwia poruszania się po tekście. Również wprowadzenie dodatków czyni pracę bardziej przejrzystą.

Moje krytyczne uwagi przedstawiam poniżej.

Uwagi ogólne

Problematyka odkrywania wiedzy na potrzeby rozwijania ontologii jest bardzo intensywnie badana od początku wieku. Niewątpliwie nie jest łatwym zadaniem prześledzenie wszystkich trendów badawczych i ich ocena. Doktorant poświęcił Rozdział 3 przeglądowi literatury w tym zakresie, jednak mając na względzie mnogość różnych podejść do zagadnienia moim zdaniem rozdział ten jest zdecydowanie za krótki i pomija kilka istotnych kierunków badań. Do najważniejszych należałoby zaliczyć:

1. Uczenie ontologii z korpusów tekstowych
2. Uczenie ontologii ze schematów bazodanowych, i/lub z diagramów ER w języku UML.

Punkt 3.1 rozprawy nawiązuje, co prawda, do uczenia ontologii z korpusów tekstowych, jest jednak zdecydowanie niekompletny. W szczególności doktorant nie wspomina prac grupy niemieckiej z uniwersytetów Karlsruhe i Koblenz (Meadche, Staab, Hotho i inni), która jako jedna z pierwszych realizowała badania nad metodami „*knowledge-poor*" bazującymi głównie na eksploracji danych tekstowych. Wymieniam tu tylko kilka ich prac, które moim zdaniem zasługują na omówienie:

1. Maedche, A., & Staab, S. (2001). Ontology learning for the semantic web. IEEE Intelligent systems, 16(2), 72-79 (ponad 2600 cytowań)
2. Maedche, A., & Staab, S. (2000, August). The text-to-onto ontology learning environment. In Software Demonstration at ICCS-2000-Eight International Conference on Conceptual Structures (Vol. 38). sn.
3. Maedche, Alexander, and Steffen Staab. "Ontology learning." Handbook on ontologies. Springer, Berlin, Heidelberg, 2004. 173-190.

Może nieco bardziej z boku (przynajmniej na pozór) leżą prace odkrywania wiedzy ze schematów i diagramów baz danych, jednak i w tym przypadku trochę mi brakuje dyskusji z takimi pracami, jak poniżej, w szczególności praca (c)

a. Hakimpour, F., & Geppert, A. (2001, October). Resolving semantic heterogeneity in schema integration. In Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001 (pp. 297-308). ACM.

b. Li, M., Du, X. Y., & Wang, S. (2005, August). Learning ontology from relational database. In Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on (Vol. 6, pp. 3410-3415). IEEE.

c. An, Y., Mylopoulos, J., & Borgida, A. (2006, July). Building semantic mappings from databases to ontologies. In AAAI (pp. 1557-1566).

Poza tym, w nawiązaniu do historii na pewno należałoby zwrócić uwagę na prace Guarino. Rozprawa nie leży, co prawda, w nurcie badań nad metodami bazującymi na analizie korpusów tekstowych, czy też odkrywania wiedzy ze schematów baz danych, jednak istotnie zyskałby rozdział badań literaturowych (a więc Rozdział 3), gdzie główne kierunki badań nad uczeniem ontologii są omówione.

Uwagi szczegółowe

1. Brak cytowań niektórych pozycji literaturowych (ja znalazłem [3], [6]);
2. Str. 17 odniesienie do pojęcia ram – warto byłoby odesłać do oryginału, tj do prac Minsky'ego, tym bardzie, że cytowana praca [19] to zauważa.
3. W p. 3.3 (Concept learning), warto byłoby nawiązać do odkrywania pojęć z tekstu;
4. W równaniu 4.2 (i innych) zbędny kwantyfikator ogólny,
5. Str. 23 pojęcie *very large number* jest nieścisłe; jeżeli wystarcza $M > max\{n,m\}$, to tak należałoby M zdefiniować.
6. Str. 31 equalities => equations
7. Str. 31 Równanie 5.3 nie jest definicją. Proponuję, aby drugą równoważność zamienić na logiczne *and*.
8. Str. 31: zdanie

   The $\prec$ relation is a partial order and thus forms the *concept lattice* of the context (G,M,I).

   jest nieścisłe, ponieważ, aby para $(U, \prec)$ była kratą, potrzebny jest dodatkowo warunek, że dla każdych 2ch x, y istnieje kres górny i kres dolny.

Większość moich uwag ma charakter redakcyjny, dlatego nie zmieniają one mojej wysokiej ocenę rozprawy.

6. Podsumowanie

Praca ma oryginalny charakter. Wkład autora jest znaczący. Jest to bardzo ciekawa propozycja w zakresie uczenia ontologii w oparciu o zasoby danych powiązanych.. Autor skutecznie łączy w pracy elementy teoretyczne z charakterem praktycznym zrealizowanych badań. Dlatego uważam, że opiniowana praca jest bardzo dobrym opracowaniem. Moim zdaniem spełnia ona z nadmiarem wymagania zawarte w obowiązujących przepisach dotyczących rozpraw doktorskich, wnoszę zatem o dopuszczenie mgra inż. Jędrzeja Potońca do publicznej obrony.