POLITECHNIKA POZNAŃSKA

WYDZIAŁ INFORMATYKI I TELEKOMUNIKACJI
Instytut Informatyki

Streszczenie rozprawy doktorskiej

AUTOMATING COMPETENCY QUESTIONS HANDLING IN ONTOLOGY DEVELOPMENT PROCESSES

AUTOMATYZACJA WYKORZYSTANIA PYTAŃ KOMPETENCYJNYCH W WYWARZANIU ONTOLOGII

Dawid Wiśniewski

Promotor
dr hab. inż. Agnieszka Ławrynowicz, prof. PP
Promotor pomocniczy
dr inż. Jędrzej Potoniec

POZNAŃ 2022

Streszczenie

Niniejsza rozprawa dotyczy problemów automatyzacji obsługi pytań kompetencyjnych w procesach wytwarzania ontologii. Ontologie są szeroko stosowane do opisu zadanej dziedziny w formie zbioru formalnie określonych pojęć i relacji między nimi [12, 6]. Pomagają one rozwiązywać zadania, wśród których wyróżnić można: automatyczne odpowiadanie na pytania, integrację informacji, ekstrakcję informacji czy ujednoznacznianie tekstu. Obecnie inżynierowie najczęściej tworzą je przy użyciu języka Web Ontology Language (OWL) [1], który bazuje na logikach deskrypcyjnych [2]. Wykorzystanie języków modelowania bazujących na logice formalnej pozwala na zastosowanie mechanizmów wnioskowania, dzięki którym istnieje możliwość odkrycia nowej, niejawnie zdefiniowanej w danej ontologii, wiedzy. Ta zdolność czyni z ontologii potężne narzędzia, które naśladują ludzki sposób rozumowania.

Wytwarzanie ontologii stanowi wyzwanie dla inżynierów. Dziedziny, które są przez nich opisywane, nierzadko są obszerne i obejmują setki lub tysiące klas i relacji wymagających sformalizowania. Z tego powodu, kluczowym wydaje się być wsparcie inżynierów w procesie identyfikacji bytów do zamodelowania i pomoc w ocenie kompletności formalizowanego słownictwa.

Co więcej, zdolność ontologii do wykorzystania mechanizmów wnioskowania powoduje, że bardzo istotną kwestią staje się zapewnienie, aby logiczne konsekwencje skonstruowanej ontologii były poprawne. Z tego powodu ważnym aspektem staje się zapewnienie inżynierom sposobów kontrolowania jakości budowanej reprezentacji, ponieważ często mają oni trudności z przewidywaniem logicznych konsekwencji modelowanej wiedzy.

Wiele metodyk z obszaru inżynierii ontologii wykorzystuje pojęcie pytań kompetencyjnych, które formułowane są w języku naturalnym i na które kompletna i poprawna ontologia powinna być w stanie poprawnie odpowiedzieć [14, 15, 5, 13]. Pytania kompetencyjne służą jako źródło słownictwa, które inżynierowie powinni zamodelować w ontologii. Podczas właściwego modelowania, inżynierowie formalizują pytania kompetencyjne przy użyciu odpowiedniego języka zapytań (np. SPARQL-OWL [7]), aby zweryfikować jakość ontologii. Poprzez obserwację, na które z pytań nie można udzielić odpowiedzi z uwagi na niekompletne słownictwo bądź na które pytania ontologia zwraca niepoprawne odpowiedzi, inżynierowie dowiadują się, które obszary modelowanej wiedzy wymagają poprawy.

Automatyzacja wykrywania słownictwa z pytań kompetencyjnych

Dotychczas obsługa pytań kompetencyjnych była głównie manualna, dlatego w niniejszej rozprawie zaproponowano dwa obszary automatyzacji ich wykorzystania [16]. Pierwszy z nich stanowi zadanie wykorzystania pytań kompetencyjnych jako źródła słownictwa, które powinno zostać zamodelowane w ontologii. Przedstawiono dwie zautomatyzowane metody dostarczające sugestie klas, instancji i właściwości do zamodelowania na podstawie zadanego zestawu pytań kompetencyjnych. Dla przykładowego pytania kompetencyjnego: Czy Microsoft Windows obsługuje pliki PDF?, aby ontologia była w stanie odpowiedzieć na pytanie, zarówno Micorosoft Windows, pliki PDF jak i po-

jęcie obsługiwania muszą zostać zamodelowane w ontologii. Celem obu zaproponowanych metod jest wykrycie takich fraz oraz zasugerowanie jak mogą być one reprezentowane w ontologii.

Pierwsza z metod oparta jest na uczeniu maszynowym i stanowi nadzorowany model warunkowych pól losowych [17]. Na podstawie zbioru treningowego zawierającego przykłady fraz oznaczonych przez eksperta jako słownictwo do zamodelowania w ontologii, zaproponowano zbiór cech, które można wydobyć z każdego tokenu (np. słowa). Cechy te, wśród których znaleźć można, między innymi, etykietę wydobytą z drzewa zależnościowego, etykietę części mowy, informację o pojawieniu się wielkiej litery w słowie, pozwalają stworzyć taką reprezentację pytania kompetencyjnego, dzięki której automatyzacja wykrywania fraz do zamodelowania jest możliwa i osiąga dobrą jakość.

Druga z metod oparta jest na ręcznie przygotowanych regułach wyrażonych w formie sekwencji znaczników części mowy [21]. Zaproponowano własny język opisu takich sekwencji i ręcznie skonstruowano zestawy reguł do wykrywania fraz przeznaczonych do zamodelowania w ontologii. Ponieważ różne reguły mogą dopasowywać się do tego samego fragmentu lub nakładających się fragmentów tekstu, zaproponowano mechanizm rozwiązywania nałożeń.

W niniejszej pracy porównano oba podejścia weryfikując ich jakość na wspólnym zbiorze testowym, a także przeanalizowano błędy popełniane przez implementacje każdej z metod. Porównanie to ukazało przewagę systemu regułowego, który generuje sugestie wyższej jakości i uogólnia się również do scenariusza wykrywania słownictwa ze zdań oznajmujących. Otrzymane wyniki dowodzą, iż możliwe jest wskazywanie słownictwa do zamodelowania w ontologii bez udziału człowieka, zachowując przy tym dobrą jakość wskazanych fraz.

Automatyzacja procesu tłumaczenia pytań kompetencyjnych do postaci zapytań SPARQL-OWL

Drugi z obszarów automatyzacji obejmuje proces formalizowania pytań kompetencyjnych do postaci języka zapytań SPARQL-OWL. W niniejszej rozprawie przedstawiono potrzebę wykorzystania języka SPARQL-OWL, a następnie zaproponowano metodę, która rekomenduje formy zapytań SPARQL-OWL dla zestawu pytań kompetencyjnych przy udziale wytwarzanej ontologii [18]. Język SPARQL-OWL to rodzaj języka SPARQL, za pomocą którego możliwe jest odpytywanie ontologii wykorzystując reżim wnioskowania o nazwie OWL 2 Direct Semantics Entailment Regime [7]. Dzięki niemu można uzyskać dostęp nie tylko do wiedzy wprost zamodelowanej w ontologii, ale również takiej, którą da się wywnioskować ze zbioru aksjomatów zawartych w ontologii.

Zaproponowana metoda oparta jest na szablonach i składa się z potoku 6 etapów przetwarzania:

1. oznaczania w pytaniu kompetencyjnym słownictwa, które występuje, bądź może występować w innej formie językowej w ontologii,
2. ekstrakcji wzorca pytania kompetencyjnego,
3. odniesienia wzorca pytania kompetencyjnego do jednego ze znanych wzorców, dla których forma zapytania SPARQL-OWL jest znana,
4. wyboru odpowiednich form zapytania SPARQL-OWL,
5. dowiązania słownictwa oznaczonego w pytaniu kompetencyjnym do encji w ontologii,
6. wypełnienia form zapytań SPARQL-OWL adresami IRI dowiązanych encji.

W rozprawie tej przeprowadzono analizę jakości narzędzia implementującego metodę na niewidzianym wcześniej zestawie pytań kompetencyjnych i ontologii. Otrzymane wyniki pozwalają sądzić, iż opracowana metoda potrafi się uogólnić na nowe, nieobserwowane wcześniej ontologie i sposoby formułowania pytań kompetencyjnych. W 46 spośród 62 przypadków testowych uzyskano oczekiwany wynik bez interwencji człowieka. Co więcej, w 34 na 37 przypadków metoda ta wykryła brak wymaganego słownictwa w ontologii.

Przedstawiono również sposób integracji tej metody z istniejącym podejściem do rozwijania ontologii wykorzystującym testy [3, 9]. W ramach wspomnianej integracji, obok tłumaczenia właściwych pytań kompetencyjnych na język zapytań SPARQL-OWL, weryfikowane są również presupozycje powiązane z pytaniami. Pytania (np. pytania kompetencyjne) zawierają ukryte założenia, zwane presupozycjami, które muszą być spełnione, aby uzyskać oczekiwane odpowiedzi [11, 4]. Przykładowe pytanie: Czy student ukończył już pracę dyplomową zakłada, że student był w trakcie przygotowywania swojej pracy dyplomowej.

Zbiór pytań kompetencyjnych i ich formalizacji w języku SPARQL-OWL

Zaproponowana metoda translacji pytań kompetencyjnych do formy zapytań SPARQL-OWL skonstruowana jest na podstawie wniosków zebranych podczas analizy zbioru przykładów pytań kompetencyjnych sformalizowanych w postaci zapytań SPARQL-OWL. Ponieważ nie istniał dotychczas żaden zbiór danych zawierający tego typu translacje, w ramach rozprawy zebrano zbiór 234 pytań kompetencyjnych zdefiniowanych dla pięciu ontologii i ręcznie przygotowano zapytania dla 131 pytań [20, 10]. W rozprawie przeanalizowano właściwości tego zbioru i zidentyfikowano regularności wśród pytań, zapytań i relacji pomiędzy obiema formami. Pokazano, że wśród zebranych pytań kompetencyjnych występuje 106 niezależnych od dziedziny wzorców pytań kompetencyjnych, spośród których część z nich jest współdzielona między pytaniami postawionymi dla różnych ontologii. Skonstruowano również niezależne od dziedziny sygnatury SPARQL-OWL, aby pokazać, że istnieją również powtarzające się wzorce wśród zapytań. Zaobserwowane regularności pomiędzy wzorcami pytań kompetencyjnych i sygnaturami zapytań SPARQL-OWL stanowiły motywację do skonstruowania metody tłumaczącej pytania kompetencyjne na zapytania SPARQL-OWL.

Ponieważ zebrany zbiór danych cechował się niewielkimi rozmiarami i nie zawierał reprezentacji wielu możliwych form pytań kompetencyjnych i zapytań, w niniejszej pracy zaproponowano również metodę automatycznego generowania par wzorców pytań kompetencyjnych i szablonów SPARQL-OWL ze zbioru wzorców aksjomatów [19]. Wzorce i szablony te mogą być następnie automatycznie uzupełniane słownictwem dziedzinowym, aby stworzyć właściwe pytania kompetencyjne i zapytania SPARQL-OWL dla zadanej ontologii.

Wykorzystano częste wzorce aksjomatów wyodrębnione z serwisu BioPortal [8], aby skonstruować syntetyczny zbiór danych składający się z niespełna 78 tysięcy wzorców pytań kompetencyjnych w relacji do 575 szablonów zapytań. Ponieważ wzorce aksjomatów reprezentują częste formy wykorzystywane do budowania aksjomatów, efekt działania metody pokrywa najpopularniejsze decyzje dotyczące zarówno sposobów modelowania wiedzy jak i formy pytań. Wykorzystanie automatycznie wygenerowanych pytań kompetencyjnych i szablonów SPARQL-OWL pozwoliło na uzyskanie znacznej poprawy jakości tłumaczenia osiąganego przez implementację metody tłumaczącej pytania kompetencyjne na język SPARQL-OWL.

Przeprowadzone w niniejszej rozprawie badania dowodzą, że możliwym jest zautomatyzowanie obsługi wykorzystania pytań kompetencyjnych w obszarach wykrywania słownictwa, które należy zamodelować w ontologii, a także tłumaczenia pytań kompetencyjnych do postaci zapytań SPARQL-OWL. Zaproponowane metody integrują się z istniejącymi metodykami wytwarzania on-

tologii, dzięki czemu możliwym jest odciążenie inżynierów i zredukowanie czasu potrzebnego na wytworzenie ontologii.

Dostępność implementacji i zbiorów danych

Wszystkie opracowane w niniejszej rozprawie metody i zbiory danych dostępne są publicznie w serwisie GitHub.

Zbiór danych 234 pytań kompetencyjnych i zapytań SPARQL-OWL znajduje się w repozytorium CQ2SPARQLOWL[1] i udostępniony jest na licencji Creative Commons Attribution 3.0 Unported License.

Syntetyczny zbiór danych składający się z niespełna 78 tysięcy wzorców pytań kompetencyjnych i 575 szablonów zapytań dostępny jest w repozytorium BigCQ[2]. Implementacje metody wykrywania słownictwa opartej na warunkowych polach losowych[3], metody wykrywania słownictwa opartej na regułach[4], jak i metody tłumaczącej pytania kompetencyjne na zapytania SPARQL-OWL[5] również dostępne są w odpowiadających im repozytoriach.

[1] `https://github.com/CQ2SPARQLOWL`
[2] `https://github.com/dwisniewski/BigCQ`
[3] `https://github.com/dwisniewski/CRFBasedGlossaryOfTermsExtraction`
[4] `https://github.com/reqtagger/ReqTagger`
[5] `https://github.com/dwisniewski/SeeQuery`

Literatura

[1] Grigoris Antoniou and Frank van Harmelen. Web ontology language: OWL. In Steffen Staab and Rudi Studer, editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 67–92. Springer, 2004.

[2] Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, USA, 2003.

[3] Kieren Davies, C. Maria Keet, and Agnieszka Lawrynowicz. More effective ontology authoring with test-driven development and the TDDonto2 tool. *International Journal on Artificial Intelligence Tools*, 28(7):1950023:1–1950023:25, 2019.

[4] Matt Dennis, Kees van Deemter, Daniele Dell'Aglio, and Jeff Z. Pan. Computing authoring tests from competency questions: Experimental validation. In Claudia d'Amato, Miriam Fernández, Valentina A. M. Tamma, Freddy Lécué, Philippe Cudré-Mauroux, Juan F. Sequeda, Christoph Lange, and Jeff Heflin, editors, *The Semantic Web - ISWC 2017 - 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part I*, volume 10587 of *Lecture Notes in Computer Science*, pages 243–259. Springer, 2017.

[5] Mariano Fernandez-Lopez, Asuncion Gomez-Perez, and Natalia Juristo. Methontology: from ontological art towards ontological engineering. In *Proceedings of the AAAI97 Spring Symposium*, pages 33–40, Stanford, USA, March 1997.

[6] Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, 1993.

[7] Ilianna Kollia, Birte Glimm, and Ian Horrocks. SPARQL query answering over OWL ontologies. In Grigoris Antoniou, Marko Grobelnik, Elena Paslaru Bontas Simperl, Bijan Parsia, Dimitris Plexousakis, Pieter De Leenheer, and Jeff Z. Pan, editors, *The Semantic Web: Research and Applications - 8th Extended Semantic Web Conference, ESWC 2011, Heraklion, Crete, Greece, May 29-June 2, 2011, Proceedings, Part I*, volume 6643 of *Lecture Notes in Computer Science*, pages 382–396. Springer, 2011.

[8] Agnieszka Lawrynowicz, Jedrzej Potoniec, Michal Robaczyk, and Tania Tudorache. Discovery of emerging design patterns in ontologies using tree mining. *Semantic Web*, 9(4):517–544, 2018.

[9] Jedrzej Potoniec, Dawid Wisniewski, and Agnieszka Ławrynowicz. Incorporating presuppositions of competency questions into test-driven development of ontologies. In *SEKE 2021 : Proceedings of the 33rd International Conference on Software Engineering and Knowledge Engineering*, pages 437–440. KSI Research Inc. and Knowledge Systems Institute Graduate School, 2021.

[10] Jedrzej Potoniec, Dawid Wiśniewski, Agnieszka Ławrynowicz, and C. Maria Keet. Dataset of ontology competency questions to SPARQL-OWL queries translations. *Data in Brief*, 29:105098, 2020.

[11] Yuan Ren, Artemis Parvizi, Chris Mellish, Jeff Z. Pan, Kees van Deemter, and Robert Stevens. Towards competency question-driven ontology authoring. In Valentina Presutti, Claudia d'Amato,

Fabien Gandon, Mathieu d'Aquin, Steffen Staab, and Anna Tordai, editors, *The Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings*, volume 8465 of *Lecture Notes in Computer Science*, pages 752–767. Springer, 2014.

[12] Barry Smith and Christopher A. Welty. FOIS introduction: Ontology - towards a new synthesis. In *2nd International Conference on Formal Ontology in Information Systems, FOIS 2001, Ogunquit, Maine, USA, October 17-19, 2001, Proceedings*, pages iii–ix. ACM, 2001.

[13] Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez, and Mariano Fernández-López. The NeOn methodology framework: A scenario-based methodology for ontology development. *Appl. Ontology*, 10(2):107–145, 2015.

[14] Mike Uschold and Michael Gruninger. Ontologies: principles, methods and applications. *Knowledge Engineering Review*, 11(2):93–136, 1996.

[15] Mike Uschold and Martin King. Towards a methodology for building ontologies. In *In Workshop on Basic Ontological Issues in Knowledge Sharing, held in conjunction with IJCAI-95*, 1995.

[16] Dawid Wisniewski. Automatic translation of competency questions into SPARQL-OWL queries. In Pierre-Antoine Champin, Fabien Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis, editors, *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pages 855–859. ACM, 2018.

[17] Dawid Wisniewski and Agnieszka Lawrynowicz. A tagger for glossary of terms extraction from ontology competency questions. In Pascal Hitzler, Sabrina Kirrane, Olaf Hartig, Victor de Boer, Maria-Esther Vidal, Maria Maleshkova, Stefan Schlobach, Karl Hammar, Nelia Lasierra, Steffen Stadtmüller, Katja Hose, and Ruben Verborgh, editors, *The Semantic Web: ESWC 2019 Satellite Events - ESWC 2019 Satellite Events, Portorož, Slovenia, June 2-6, 2019, Revised Selected Papers*, volume 11762 of *Lecture Notes in Computer Science*, pages 181–185. Springer, 2019.

[18] Dawid Wisniewski, Jedrzej Potoniec, and Agnieszka Lawrynowicz. SeeQuery: An automatic method for recommending translations of ontology competency questions into SPARQL-OWL. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 2119–2128, New York, NY, USA, 2021. Association for Computing Machinery.

[19] Dawid Wisniewski, Jedrzej Potoniec, and Agnieszka Lawrynowicz. Bigcq: Generating a synthetic set of competency questions formalized into SPARQL-OWL (student abstract). In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 13079–13080. AAAI Press, 2022.

[20] Dawid Wiśniewski, Jedrzej Potoniec, Agnieszka Ławrynowicz, and C. Maria Keet. Analysis of ontology competency questions and their formalizations in SPARQL-OWL. *Journal of Web Semantics*, 59:100534, 2019.

[21] Dawid Wiśniewski, Jędrzej Potoniec, and Agnieszka Ławrynowicz. Reqtagger: A rule-based tagger for automatic glossary of terms extraction from ontology requirements. *Foundations of Computing and Decision Sciences*, 47(1):65–86, 2022.