Kraków, 14 czerwiec 2019

Prof. dr hab. inż. KrzysztofyZieliński

Katedra Informatyki

Wydział Informatyki Elektroniki i Telekomunikacji

AGH

Recenzje rozprawy doktorskiej mgr inż. Adama Łuszpaja pt.

„Zdecentralizowana infrastruktura do integracji wiedzy w Semantic Web"

1. Wprowadzenie

Zagadnienia pozyskiwania i integracji wiedzy znajdują się w centrum uwagi współczesnej informatyki. Dotyczy to w szczególności wiedzy gromadzonej w sieci Web, w której w zdecydowanej większości informacja jest wyrażona w języku naturalnym i przeznaczona dla człowieka. Stanowi to zasadniczą barierę wykorzystania tej wiedzy przez inteligentne aplikacje, które powinny rozumieć dane. W celu jej przełamania kilka lat temu zaproponowano Semantic Web, który wykorzystuje RDF do wyrażania wiedzy oraz ontologie w procesach jej strukturalizacji i przewarzania. Pomimo, że Semantic Web stanowił milowy krok jednakże nie spełnił do końca istniejących oczekiwań. Zdaniem doktoranta stało się tak z powodu braku wystraczająco efektywnej infrastruktury dostępu do wiedzy i modelu skutecznego wykorzystania zbiorów Web of Data.

Stanowi to motywację dla doktoranta do opracowania koncepcji zdecentralizowanej infrastruktury dostępu i integracji wiedzy w Semantic Web. Wykazanie wykonalności takiej infrastruktury przy spełnieniu postulatów efektywnego pozyskiwania wiedzy stanowi tezę rozprawy i przedmiot pracy.

Przedstawioną tematykę rozprawy oraz sformułowane zadanie badacze uważam za ważne i aktualne zarówno z teoretycznego i praktycznego punktu widzenia.

2. Omówienie zawartości rozprawy

Praca składa się z 6 rozdziałów z których pierwszy stanowi wstęp. Rozdział 2 zawiera dosyć obszerne wprowadzenie do tematyki dostępu do wiedzy w Semantic Web. Na wstępie tego rozdziału w sposób bardzo uporządkowany omówiono architekturę, użycie, problemy i niedostatki systemów WWW, Semantic Web oraz Linked Data. Wskazano, że podstawowym oczekiwaniem w stosunku do Semantic Web jest aktywne zachowanie sieci w procesie pozyskiwania wiedzy co oznacza, że odpowiedź na zapytanie nie powinna być udzielana na podstawie jednego źródła danych lecz możliwie pełnej wiedzy potencjalnie wykorzystującej wiele źródeł. Realizacja tego procesu, jak stwierdza autor napotyka jednakże na dwie bariery: architektoniczną – brak połączeń bądź standardów jednolitego dostępu do wiedzy oraz semantyczną – wykorzystanie w tych samych obszarach tematycznych różnych ontologii. W dalszym ciągu tego rozdziału autor prowadzi rozważania jak można przełamać te bariery. W tym kontekście doktorant przedstawia ewolucję paradygmatów dostępu do wiedzy i wytycza kierunek

własnych prac formułując zarys koncepcji budowy infrastruktury pozyskiwania wiedzy o nazwie ActiveDiscovery.

Rozdział 3 przedstawia szczegółowo koncepcję infrastruktury ActiveDiscovery spełniającej sformułowane postulaty efektywnościowe. Punktem wyjścia proponowanego rozwiązania jest diagnoza obecnie realizowanego sposobu dostępu do wiedzy i realizacji zapytań w Semantic Web i Linked Data. Prowadzi ona do bardzo ważnego założenia, iż niezależnie od reprezentowanego zasobu każdy identyfikator URI może być traktowany jako reprezentacja lokalizacji sieciowej zawierającej dane powiązane z przedmiotem identyfikacji. Następnie zostały przedstawione elementy infrastruktury ActiveDiscovery: adnotacje, węzły oraz kluczowe usługi. W kolejnym kroku przedstawiono protokół realizacji zadania pozyskiwania wiedzy w wersji podstawowej, który na kolejnym etapie zostaje uzupełniony do wersji integracyjnej uwzględniającej odwzorowania pomiędzy ontologiami. W dalszym ciągu rozważań zostały podane procedury przetwarzania zapytań wraz z wykazaniem ich poprawności. Rozważania te prowadzone są bardzo konsekwentnie oraz domknięte dyskusją złożoności obliczeniowej oraz omówieniem zagadnień optymalizacji zapytań w architekturze federacyjnej.

Rozdział 4 zawiera ewaluację eksperymentalną infrastruktury ActiveDiscovery. Przedmiotem badań była procedura rozproszonego wykonania zapytań zarówno w wersji nie korzystającej z odwzorowań między ontologiami, jak również z wykorzystaniem tych odwzorowań. Doktorant konsekwentnie stara się pokazać empirycznie , że zgodnie z przyjętą tezą rozprawy, pozyskanie wiedzy, w rozumieniu Semantic Web, przez klienta końcowego nie wymaga apriorycznej informacji na temat lokalizacji tej wiedzy. Takie podejście wymagało od doktoranta implementacji prototypowej implementacji infrastruktury ActiveDiscovery, opracowania rzeczywistych scenariuszy użytkowych zapytań oraz danych na które składają się: ontologie OWL, asercyjne zbiory RDF, oraz odwzorowania pomiędzy tymi ontologiami.

Przyjęta metodologia ewaluacji obejmuje zarówno ocenę jakościową jak i ilościową. Ewaluacja jakościowa sprowadza się do wykazania, że zbiór wyników zwracanych przez ActiveDiscovery jest nadzbiorem sumy wyników zwracanych przez zapytania skierowane bezpośrednio do odpowiednich źródeł. Ewaluacja ilościowa jest zorientowana na wykazanie skalowalności opracowanej infrastruktury. Dla tak sformułowanego zadania ewaluacji zasadnicze znaczenie posiada dobór zapytań testowych. W tym celu doktorant wykorzystuje wyniki dotychczas opublikowanych prac adresujących to zagadnienie i wybiera odpowiednio elementy stałe i zmienne zapytań. Takie podejście jest w pełni uzasadnione i daje możliwość porównania uzyskanych wyników z tymi opisanymi w publikacjach. Rozdział kończy opis środowiska ewaluacyjnego ActiveDiscovery.

W Rozdziale 5 przedstawiono wyniki ewaluacji środowiska ActiveDiscovery wykorzystując ontologie, dane i zapytania tematycznie związane z handlem elektronicznym. Obejmują one zarówno ewaluację jakościową jak i ilościową. Ewaluacja jakościowa potwierdziła, że wykonanie zapytań niezlokalizowanych w infrastrukturze ActiveDiscovery daje lepsze wyniki niż zapytania skierowane do źródeł wiedzy nie wykorzystującej tej infrastruktury. Potwierdza to zasadność tezy rozprawy oraz pozytywnie weryfikuje projektowane własności infrastruktury ActiveDiscovery. Z punktu widzenia rozważań nad skalowalnością środowiska ActiveDiscovery istotne jest określenie jak zmienia się czas wykonania zapytania wraz ze wzrostem rozmiaru zbiorów danych RDF dla różnych warunków

parametryzujących tą funkcję. Dotyczą one: wielkości zapytania mierzonego liczbą wzorców trójkowych w BGP, selektywności zapytania kontrolowanej za pomocą filtrowania i umieszczania instancji w zapytaniu oraz liczby źródeł danych. Przyjęte podejście prowadzi do wyróżnienia ośmiu typów zapytań dla których prowadzona była ewaluacja. Należy podkreślić, że ewaluację przeprowadzono zarówno na danych generowanych jak i na kilku przykładach źródeł danych i ontologii rzeczywiście występujących w Sematic Web takich jak DBpedia, FactForge oraz Schema.org. Uzyskane wyniki potwierdziły praktycznie liniową skalowalność opracowanej infrastruktury względem badanych parametrów. Jednocześnie stwierdzono, że rozmiar zapytania wydaje się być kluczowym parametrem decydującym o koszcie.

Rozdział 6 zawiera podsumowanie pracy, najważniejszych jej osiągnięć oraz przedstawienie proponowanych kierunków dalszych prac.

Pracę kończy spis literatury obejmujący 174 pozycje, z których zdecydowana większość jest dostępna w sieci Internet.

3. Ocena merytoryczna

Praca zawiera ciekawą propozycje infrastruktury rozproszonej do efektywnego pozyskiwanie wiedzy w systemie Semantic Web. Jednym z głównych pomysłów jest wykorzystanie semantycznej warstwy zapytania jako klucza do odszukana źródeł danych dla sformułowania odpowiedzi. Wymaga to jednakże spełnienia założenia, że URI lokalizuje zasób informacji oraz zbudowania odpowiedniej infrastruktury zbudowanej z pięciu węzłów z których dwa stanowią magazyny wiedzy terminologicznej i przechowują ontologie. Dwa kolejne stanowią rozproszony rejestr metadanych : indeksu łączącego ontologie ze źródłami danych oraz indeksu odwzorowań pomiędzy ontologiami. Połączenie ontologii ze źródłami danych wymagało wprowadzenia elementów adnotacji dla ontologii OWL. Piąty węzeł jest mediatorem koordynującym wykonanie zapytań na wielu węzłach danych źródłowych. Wykorzystuje on dane udostępniane przez węzły indeksowe. Dla wybranej klasy zapytań skonstruowano Algorytm LOC translacji niezaadresowanego zapytania źródłowego na zapytania wynikowe adresowane do wskazywanych przez węzły indeksowe źródeł danych. Doktorant uwodnił przy tym twierdzenie o poprawności tego algorytmu. Jest to w moim przekonaniu bardzo wartościowe osiągnięcie doktoranta. Ważnym rozszerzeniem tego osiągnięcia jest opracowanie algorytmu odwzorowania pojęć pomiędzy ontologiami (Algorytm MAPLOC). Poprawność tego algorytmu także została udowodniona formalnie.

Przyjęta koncepcja rozwiązania jest logicznie spójna a jej poprawność w sensie algorytmicznym wykazano na drodze teoretycznej natomiast efektywność potwierdzono przemyślanymi badaniami eksperymentalnymi.

Oprócz opracowania koncepcji infrastruktury ActiveDiscovery do osiągnięć doktoranta należy zaliczyć:

- implementację zaproponowanej infrastruktury ActiveDiscovery,
- opracowanie przemyślanej strategii badań jakościowych i ilościowych oraz odpowiednich scenariuszy,
- analizę złożoności problemu ewaluacji zapytań i optymalizacji ich przetwarzania,
- optymalizację heurystyczną rozszerzającą, usługi węzła indeksowego o dodatkowe informacje wpływające na selektywność zapytań.

Wymienione elementy sprawiają, że całość pracy można uznać za dobrze przemyślaną i precyzyjnie dokumentującą konsekwentnie realizowany proces badawczy. Doktorant umiejętnie łączy wątki teoretyczne z zagadnieniami praktycznymi oraz właściwie wspiera swoje wywody badaniami literaturowymi.

Praca jest dobrze napisana a jej klarowna struktura sprawia, że dobrze się czyta. Znalazłem tylko nieliczne usterki redakcyjne.

4. Uwagi krytyczne polemiczne

Lektura każdej rozprawy naukowej stanowi podstawę do sformułowania uwag o charakterze, krytycznym czy też polemicznym. Nie oznacza to jednakże, iż przedstawiona rozprawa doktorska jest obarczona niedociągnięciami. W rozważanym przypadku doktorant bardzo konsekwentnie zrealizował proces badawczy i bez wątpienia wykazał tezę rozprawy jaką postawił na początku procesu badawczego. Należy podkreślić, że teza ta jest precyzyjnie postawiona wobec jasnej definicji postulatów efektywnego pozyskiwania wiedzy w środowisku Semantic Web tj.: aktywności sieci, transparentności, kompletności, prymatu i neutralność ontologii oraz decentralizacji i skalowalności. Pozwala to na jednoznaczną ocenę czy infrastruktura ActiveDiscovery spełnia postawione wymagania i w kontekście przedstawionych wyników badań jest to ocena zdecydowanie pozytywna. Fakt ten skłania do postawienia pytań dotyczących praktycznego wykorzystania uzyskanych rezultatów. Odpowiedzi na te pytania można uznać jako leżące poza zakresem pracy jednakże ich sformułowanie pozwala na lepszą ocenę znaczenia przedstawionego osiągnięcia naukowego.

Efektywność infrastruktury ActiveDiscovery została przekonywująco wykazana jednakże, jej budowa a zwłaszcza stworzenie węzłów indeksujących nie jest zadaniem prostym. Wymaga pozyskania informacji o rozlokowaniu źródeł wiedzy. Powstaje zatem pytanie jak w środowisku Semantic Web automatycznie lub semi-automatycznie wygenerować infrastrukturę ActiveDiscovery. Jakie narzędzia należy opracować aby usprawnić ten proces.

Raz wygenerowana infrastruktura ActiveDiscovery ulega dezaktualizacji ponieważ w sieci powstają nowe źródła wiedzy a stare mogą ulegać destrukcji. Jest to proces dynamiczny, który wymusza potrzebę stałej aktualizacji infrastruktury ActiveDiscovery, podobny do tego, który odbywa się w systemie DNS. Powstaje w tym kontekście pytanie jak zidentyfikować zdarzenia, których wystąpienie powinno wymusić proces uaktualnienia infrastruktury ActiveDiscovery.

Badania przeprowadzone przez doktoranta dotyczyły klasy zapytań, która zgodnie z badaniami literaturowymi posiada największe znaczenie praktyczne. Jednakże otwartym pozostaje problem skutecznego rozszerzenia infrastruktury ActiveDiscovery na inne klasy zapytań.

W kontekście postawionych pytań ciekawe jest także poznanie opinii doktoranta co do szansy szerszego praktycznego wdrożenia infrastruktury ActiveDiscovery.

5. Podsumowanie

Rozprawa doktorska mgr inż. Adama Łuszpaja pt. „Zdecentralizowana infrastruktura do integracji wiedzy w Semantic Web" zawiera rozwiązanie oryginalnego aktualnego problemu badawczego, który jest ważny z punktu widzenia procesu automatycznego pozyskiwania wiedzy z sieci Web. Doktorant wykazał się bardzo dobrą znajomością tematyki sieci Semantic Web oraz właściwie zidentyfikował wzywania istniejące w tym obszarze. Sformułował ciekawe zadanie badawcze i w sposób pomysłowy je rozwiązał, pokazując przy tym, iż opracowane rozwiązanie spełnia postawione na wstępie złożenia efektywnościowe. Należy także podkreślić, że tematyka rozprawy dotyczy ważnych obszarów współczesnej informatyki.

W związku z powyższym stwierdzam, że przedstawiona rozprawa doktorska mgr inż. Adama Łuszpaja spełnia wszystkie wymagania stawiane przez ustawę o stopniach i tytułach naukowych i wnoszę o dopuszczenie doktoranta do dalszych etapów przewodu doktorskiego. Jednocześnie z uwagi na zakres tematyczny rozprawy i charakter przedstawionego osiągnięcia naukowego stwierdzam, że mgr inż. Adam Łuszpaj powinien otrzymać stopień doktora w dyscyplinie naukowej Informatyka Techniczna i Telekomunikacja.