Wrocław 15.06.2019 r.

Prof. dr hab. inż. Ngoc Thanh Nguyen
Katedra Systemów Informatycznych
Wydział Informatyki i Zarządzania
Politechnika Wrocławska

Recenzja rozprawy doktorskiej mgr inż. Adama Łuszpaja

pt.

„Zdecentralizowana infrastruktura do integracji wiedzy w Semantic Web”

Recenzja niniejsza została przygotowana na zlecenie Dziekana Wydziału Wydział Informatyki, Elektroniki i Telekomunikacji, Akademia Górniczo-Hutnicza im. Stanisława Staszica (pismo z dnia 12.04.2019).

I. Wprowadzenie

Sieć Web obecnie jest największym źródłem danych z najszybszym i najbardziej powszechnym dostępem. Dzieje się tak dzięki m.in. prostym strukturom danych w niej zawartych. Sieć semantyczna (*Semantic Web*) natomiast powstała na potrzebę przetwarzania semantycznego. Jej rozproszone i autonomiczne zasoby są wykorzystane za pomocą inteligentnych aplikacji. Jednak, co słusznie zauważył Autor rozprawy, pomiędzy warstwą semantycznego opisu danych a warstwą inteligentnych aplikacji brakuje infrastruktury pośredniej, która integrowałaby wiedzę z różnych źródeł tworząc dane wejściowe dla tych aplikacji.

To zagadnienie stanowi motywację i problem badawczy dla Autora niniejszej rozprawy doktorskiej. Jest ono jak najbardziej aktualne, wysoce praktyczne i wymaga efektywnego rozwiązania.

Jako cel pracy Autor postawił na opracowanie infrastruktury dostępu i integracji wiedzy, która wypełnia wyżej wymienioną lukę i przybliża realizacje zadania pozyskiwania wiedzy. Autor sformułował 6 postulatów efektywnego pozyskiwania wiedzy i sformułował jako teza pracy następujące stwierdzenie:

„*Możliwe jest opracowanie zdecentralizowanej infrastruktury dostępu i integracji wiedzy w Semantic Web, spełniającej określone postulaty efektywnego pozyskiwania wiedzy*”.

W mojej ocenie problem badawczy jak i cel rozprawy oraz teza zostały jasno i spójnie sformułowane.

II. Struktura rozprawy

Praca liczy 164 strony, składa się z 6 rozdziałów oraz tabel i rysunków. Jej treść można scharakteryzować następująco:

Rozdziały 1 zawiera wprowadzenie do zagadnienia oraz uzasadnienie problemu badawczego, cel i tezę pracy.

Rozdział 2 zawiera przegląd stanu wiedzy i kierunków badawczych dotyczących zagadnień dostępu do informacji w sieci Web. Autor przedstawił także ewolucję podejść do problemu integracji i pozyskiwania wiedzy w wysoce zdecentralizowanym i heterogenicznym środowisku Semantic Web.

Rozdział 3 szczegółowo przedstawia koncepcje autorskiej infrastruktury ActiveDiscovery służącej do realizacji zadania pozyskiwania wiedzy i spełniającej postulaty efektywnościowe. Punktem wyjścia jest diagnoza obecnego stanu rzeczy dotyczącego dostępu do wiedzy i realizacji zapytań w Semantic Web i Linked Data. Rozdział zawiera opis koncepcji zdecentralizowanej infrastruktury, której założenia są oparte o zasady Semantic Web. Przedstawione zostały elementy strukturalne, m.in. węzły i ich usługi, protokoły realizacji zadania pozyskiwania i integracji wiedzy, charakterystykę klas zapytań, procedury translacji w wersji podstawowej i integracyjnej, a także analiza złożoności opracowanych algorytmów, kosztu ich wykonania i technik optymalizacji zapytań.

W rozdziale 4 Autor opisuje proces ewaluacji, której celem jest eksperymentalna walidacja scenariusza pozyskiwania wiedzy zaprezentowanego w tezie pracy. W pierwszej kolejności przedstawił metodologię ewaluacji określającą strategię testowania i kryteria oceny wykonalności opracowanej infrastruktury. Następnie podana jest charakterystyka prototypowej implementacji ActiveDiscovery, zapytań reprezentujących realistyczne scenariusze użytkowe oraz danych testowych Opis środowiska ewaluacyjnego jest także załączony.

Rozdział 5 zawiera opis eksperymentów zawierających założone zapytania wraz z kontekstami ich wykonania, zarówno dla ewaluacji jakościowej jak i ilościowej. Prezentowane są wyniki uzyskiwane dla poszczególnych zapytań przy użyciu prototypowej

implementacji. Zostały przedstawione dwa przykłady prezentujące wartość dodaną dla rzeczywistych danych i usług w DBpedii i FactForge z ograniczeniami dotyczącymi konieczności uzupełnienia o elementy wymagane w opracowanej infrastrukturze ActiveDiscovery.

Rozdział 6 zawiera podsumowanie pracy i kierunki dalszych badań.

W pracy zacytowano 174 prace, wśród nich nie ma żadnych publikacji z udziałem Autora rozprawy.

III. Omówienie oryginalnego wkładu naukowego rozprawy

Poniżej opisuję najważniejsze wyniki rozprawy osiągnięte przez Doktoranta oraz komentuję poszczególne elementy. Uwagi krytyczne do tych elementów będą pisane pochyłą pisownią.

- **Zdefiniowanie elementów strukturalnych infrastruktury ActiveDiscovery**. Autor skonstruował graf z węzłami, które stanowią ważną część tej infrastruktury. Następujace węzły zostały specyfikowane: *Ontology Repository Node* (ORN) jako repozytorium ontologiczne. Jego zadaniem jest udostępnianie ontologii zgodnie z protokołem http; węzeł *Data Repository Node* (DRN) jako repozytorium danych RDF i eksponuje punkt dostępowy SPARQL do swojej zawartości; węzeł *Indexing Service Node* (ISN) przechowujący powiązania pomiędzy TBoxami a odpowiadającymi im ABoxami oraz węzeł *Querying Service Node* (QSN) świadczący usługę realizacji niezaadresowanego zapytania (QueryService), która jest w istocie implementacją zadania efektywnego pozyskiwania wiedzy. Autor określił jasno powiązania pomiędzy tymi elementami i ta koncepcja wygląda na racjonalną oraz logiczną. *Jednak opis ten jest na bardzo wysokim poziomie ogólności i wymaga ukonkretnienia struktur tych elementów. Np. jak ontologie są składane do węzła ORN, w jakiej kolejności, jak wygląda sprawa mapowania pomiędzy nimi? Jakie są rodzaje powiązań pomiędzy TBoxami i ABoxami?*
- **Opracowanie protokołów dostępu do wiedzy**. Wyszczególnione są 2 procesy: rejestracja wiedzy i jej pozyskanie. Odnośnie rejestracji wiedzy, proces ten jest opisany za pomocą diagramu 3.3, który wygląda w sensowny sposób. Diagram ten ilustruje sekwencje wymiany komunikatów pomiędzy węzłami, które zmierzają do dodania odpowiednich wpisów indeksowych na zdeterminowanym przez ontologie węźle ISN. *Opis tej części pracy oddaje jednak za mało szczegółów. Nie wiadomo jak niektóre komunikaty wyglądają konkretnie. Np. komunikat ekstrakcji TBox czy komunikat*

dodajWpisyIndeksowe i zapisanie powiązania pomiędzy zbiorem danych i ontologią. *Te komunikaty w istocie są złożonymi procedurami, które dobrze byłoby opisać.* Odnośnie procesu pozyskiwania wiedzy, diagram 3.4 jasno opisuje koncepcję. *Jednak podobnie jak wyżej, czytelnik oczekuje na podanie więcej konkretów, np. jak wygląda proces translacji zapytania czy delegacji zapytania.*

- **Opracowanie metody integracji wiedzy**. Integracja na poziomie danych w ramach *Linked Data* opiera się na tworzeniu linków będących standardowymi trójkami RDF, które łączą rozproszone zbiory danych. Na poziomem integracji semantycznej proces ten jest znacznie bardziej utrudniony poprzez różnorodność w strukturze jak i znaczeniu elementów ontologii (pojęcia, relacje), które modelują niekiedy te same obszary dziedzinowe. W rozprawie Autor zakłada, że odwzorowanie ma charakter luźnego, autonomicznego powiązania pomiędzy elementami ontologii sieciowych. *To założenie jest niezrozumiałe dla mnie.* Autor zdefiniował dodatkowy mechanizm, analogiczny do opisanego indeksu wiążącego warstwę ontologii z warstwą wiedzy instancyjnej. Schemat na rysunku 3.7 pokazuje proces pozyskiwania wiedzy z uwzględnieniem odwzorowań między ontologiami, który zawiera etap odpytania węzłów MSN dla poszczególnych ontologii zapytania w poszukiwaniu ontologii zawierających skojarzone pojęcia. W rezultacie pozyskania odwzorowań usługa QueryService może odnosić się do większej liczby ontologii. Dzięki temu możliwe jest rozszerzenie spektrum potencjalnych węzłów danych mogących uczestniczyć w odpowiedzi na zapytanie. Pokazuje to spełnienie postulatu zmaksymalizowania liczby źródeł danych (postulat kompletności) oraz postulatu niezależności wyników od konkretnej ontologii użytkownika (postulat neutralności ontologii). *Należy jednak zwrócić uwagę, że proces integracji wiedzy, m.in. integracji ontologii, jest bardzo złożonym, nie tylko czasowo, procesem. Autor zdaje się za bardzo tym się nie przejmuje, choć to ma bardzo istotny wpływ na efektywność opracowanej infrastruktury*. *Opisana powyżej koncepcja jest logiczna i racjonalna, jednak brak jest przedstawienia konkretnych procedur realizujących te elementy.*

- **Opracowanie koncepcji realizacji zapytań**. Autor opiera się na narzędziu SPARQL. Zapytania skierowane do sieci są wykonywane przez węzeł QSN i jego usługę QueryService. Potrzebne jest zebranie informacji na temat źródeł danych. W tym celu wykorzystywane są informacje o ontologiach oraz o ich lokalizacjach i odwzorowaniach pomiędzy nimi. Następnie należy wykonać przekształcenie zapytania w oparciu o zebrane informacje. Nieadresowane zapytanie zostaje tłumaczone na zapytanie,

wyposażone w informacje o punktach dostępowych do wiedzy znaczeniowo związanej z zapytaniem wejściowym. W tym celu Autor zdefiniował pewne pojęcia jak wzorzec trójkowy, wzorzec grafowy, przedefiniowanie słowników, zapytania ABox i ich rozszerzenie semantyczne. Autor opracował algorytm LOC służący do przekształcenia zapytania, który jest wykonywany jest przez węzeł QSN w ramach usługi QueryService. Udowodnił (Twierdzenie 1), że jest on poprawny. Jest to ważny element rozprawy, świadczący o prawidłowości w postępowaniu Autora przy opracowaniu algorytmów. Dowód nie jest trywialny.

- **Opracowanie algorytmu realizacji zapytania z uwzględnieniem odwzorowań**. Autor uzupełnił podstawową infrastrukturę o elementy usługowe związane z dostarczaniem odwzorowań dla celów integracji pozyskiwanej wiedzy. Sposób wykorzystania tych elementów wpływa na kształt algorytmu przepisywania zapytania. Następnie rozszerzył algorytm LOC o wykorzystanie indeksu odwzorowań oraz samych odwzorowań do maksymalizacji potencjalnie użytecznych źródeł budujących odpowiedzi. Autorski algorytm o nazwie MAPLOC jest istotny ponieważ odwzorowania są bardzo częstymi elementami w przetwarzaniu semantycznym. Także i w tym przypadku Autor udowodnił, że opracowany algorytm jest poprawny (Twierdzenie 2).
- **Zbadanie złożoności algorytmów i optymalizacji zapytań dla opracowanej infrastruktury**. Są to elementy istotne w opracowaniu nowych języków zapytań i Autor jest świadom tej konieczności. Przedstawił dość szczegółową i prawidłową analizę tych elementów.
- **Opracowanie analizy efektywności infrastruktury ActiveDiscovery**. W pierwszej części Autor opisał metodologię ewaluacji określającą strategię testowania i kryteria oceny wykonalności opracowanej infrastruktury. Porównał także infrastrukturę ActiveDiscovery z architekturą DNS. Została opisana charakterystyka prototypowej implementacji ActiveDiscovery, zapytań reprezentujących realistyczne scenariusze użytkowe oraz danych testowych przygotowanych zgodnie z przyjętą metodologią. Autor opracował także środowisko ewaluacyjne. Następnie przeprowadził eksperymenty związane z procedurą rozproszonego wykonania zapytania, zarówno w wersji nie korzystającej z odwzorowań między ontologiami, jak również z ich wykorzystaniem. Dobór zapytań testowych jest bardzo istotny w tym etapie. Autor dokonał tego wyboru biorąc pod uwagę przyjętą strategię ewaluacji, zwłaszcza w związku z badaniem skalowalności. Wymaga ona sformułowania zapytań zróżnicowanych pod względem

rozmiaru i selektywności. Autor dobrze uzasadnił ten wybór. Wyniki eksperymentów zostały zaprezentowane w sposób czytelny i dokładny. Cel przeprowadzenia eksperymentów został osiągnięty.

- **Przeprowadzenie dowodu spełnienia zdefiniowanych 6 postulatów efektywności przez opracowaną infrastrukturę.** Autor często wspomniał o ich spełnieniu budując i testując architekturę ActiveDiscovery (np. na str. 16, 26, 52, 53) i można w zasadzie zaakceptować ten „miękki" dowód, choć jest on bardzo lakoniczny. *Pozostaje jednak pewien niedosyt odnośnie kompletności i głębokości tego dowodu. Np. postulat praktycznej kompletności stanowiący, że sieć powinna zlokalizować i wykorzystać jak największą liczbę źródeł wiedzy przydatnych do skonstruowania odpowiedzi. Trudno w pełni uznać, że zaproponowana procedura zapewnia spełnienie to kryterium. To samo z postulatem nr 4 (o prymacie ontologii). Należy zdawać sobie sprawę o dużej trudności przeprowadzenia pełnych i głębokich dowodów dla tych postulatów. Taka trudność wynika m.in. z dużej ogólnikowości w sformułowaniu tych postulatów oraz w braku głębszych specyfikacji w określeniu elementów zaproponowanej infrastruktury. Recenzowana rozprawa jednak zostawia wrażenie o zbyt powierzchownym ich traktowaniu.*

Pomimo pewnych krytycznych uwag, w świetle opisu tych osiągnięć jestem w stanie stwierdzić, że opracowana przez Autora infrastruktura o nazwie ActiveDiscovery i analiza, w tym eksperymentalna jej własności, jest oryginalna i spełnia założone cele. Opisane osiągnięcia Doktoranta świadczą o jego istotnym wkładzie do dziedziny przetwarzania semantycznego. Rozprawa doktorska mgr. inż. Adama Łuszpaja ma bardzo także duży aspekt praktyczny i dobrze wpisuje się w najnowszy, szczególnie ważny i innowacyjny nurt światowych badań w dziedzinie Semantic Web. Uzyskane wyniki świadczą o tym, że cel pracy został osiągnięty, a teza udowodniona. Praca jest dobrze i poprawnie napisana. Redakcja nie budzi zastrzeżeń.

V. <u>Konkluzje</u>

Stwierdzam, że rozprawa doktorska mgr inż. Adama Łuszpaja, stanowiąc oryginalne rozwiązanie problemu naukowego oraz wykazując głęboką wiedzę Doktoranta w dziedzinach Semantic Web i inżynierii ontologii, a także umiejętność samodzielnego prowadzenia badań naukowych, spełnia wymagania określone w Ustawie o Stopniach naukowych i Tytule naukowym. Ponadto zakres i jakość wyszczególnionych badań

przeprowadzonych przez Doktoranta w pełni świadczą o jego solidnym wkładzie do tych dziedzin.

Dodatkowo, z czysto parametrycznego punktu widzenia, dorobek Doktoranta uzyskany w czasie powstawania rozprawy składający się z 7 publikacji, w tym jeden artykuł w czasopiśmie, 3 referaty opublikowane w materiałach międzynarodowych konferencji i 3 rozdziały w wydawnictwach zwartych, można uważać za dobry.

Wobec powyższego wnioskuję o dopuszczenie rozprawy do publicznej obrony.