Інформаційні технології

УДК 004.852

Галушко Марія Олегівна
бакалавр комп'ютерних наук,
Національний технічний університет України
«Київський політехнічний інститут»
Галушко Мария Олеговна
бакалавр компьютерных наук,
Национальный технический университет Украины
«Киевский политехнический институт»
Halushko M.
Bachelor of computer science
The National Technical University of Ukraine
«Kyiv Polytechnic Institute»

ДОСЛІДЖЕННЯ ЗДІЙСНЕННЯ «РОЗУМНОГО» ПОШУКУ У RDF-СХОВИЩАХ
ИССЛЕДОВАНИЕ ОСУЩЕСТВЛЕНИЯ «УМНОГО» ПОИСКА В RDF-ХРАНИЛИЩАХ
RESEARCH OF IMPLEMENTATION "SMART" SEARCH IN RDF-STORES

Анотація: Досліджено можливість здійснення «розумного» пошуку у RDF-сховищах, таких як бібліотека наукових публікацій.

Ключові слова: RDF, OWL, онтологія, база знань, семантичний веб, SPARQL, середовище опису ресурсів, триплет.

Аннотация: Исследована возможность осуществления «умного» поиска в RDF-хранилищах, таких как библиотека научных публикаций.

Ключевые слова: RDF, OWL, онтология, база знаний, семантический веб, SPARQL, среда описания ресурсов, триплет.

Abstract: The possibility of "smart" search RDF-storage facilities such as library publications.

Keywords: RDF, OWL, ontology, base of knowledge, semantic web, SPARQL, resource description environment, triple.

Зараз значно збільшується потік інформації, з'явилася необхідність пошуку нових способів її зберігання, подання, формалізації і систематизації, а також автоматичної обробки. При вирішенні задач, в яких дані можуть мати довільні зв'язки, виникає непередбачувана кількість зв’язків в запитах, тому для вирішення такого плану задач, зараз найбільшого розповсюдження набули RDF-сховища. Вони базуються на стандартах комітету W3C для мови опису графів (RDF) та для обробки графових даних (SPARQL). RDF значить “середовище опису ресурсів”. [1, с. 204]. Це модель даних, що представляє дані простими триплетами суб’єкт – предикат – об’єкт. Триплет також відомий як «оператор» і є базовим «явищем», або його ще можна назвати стверджуваним блоком знань в RDF. Декілька операторів комбінуються разом узгоджуючись таким чином: суб’єкт, об’єкт як вузол і предикат як ребро. Таким чином, виникає структурна мережа, також відома як RDF-граф і має вигляд вузол-ребро-вузол.

RDF визначає загальну архітектуру метаданих і призначена для забезпечення сумісності метаданих за допомогою спільної семантики, структури та синтаксису. Технологія семантичної мережі передбачає розширення можливостей інтернету завдяки механізмам надання інформації чітко визначеного значення, яке дозволяє ефективно використовувати її у спільній роботі як комп'ютерів, так і людей [2, с. 36-57].

Онтологія – це загальноприйнята і загальнодоступна концептуалізація певної області знань (світу, середовища), яка містить

базис для моделювання цієї області знань і визначає шляхи для взаємодії між агентами, які використовують знання з цієї області, і, нарешті, включає домовленості про представлення теоретичних основ даної області знань.

У загальному вигляді структура онтології являє собою набір елементів чотирьох категорій:

- ***Поняття*** (представники якоїсь сутності або явища, описують групу індивідуальних сутностей, які об'єднані на підставі наявності загальних властивостей).
- ***Відношення*** (відношення IS-A; клас - підклас; відношення a-kind-of).
- ***Аксіоми*** (висловлюють ту інформацію, яка не може бути відображена в онтології за допомогою побудови ієрархії понять. Н-д, «Якщо X смертний, то X колись помре»).
- ***Окремі екземпляри*** (конкретні елементи будь-якої категорії (наприклад, екземпляром класу Студент буде Віктор) [3, с. 50-64].

Основними етапами створення онтологій є:

1. Визначення класів.
2. Створення ієрархії класів.
3. Визначення слотів (властивостей).

Наприклад, «використовує», «належить», «вирішує» (проблему), і т.д.

4. Заповнення онтології екземплярами та встановлення зв’язків між об’єктами [4, с. 210-221].

Під час виконання було проаналізовано існуючі аналоги, такі як Google Scholar, DBPedia і Scirus.

Незважаючи на те, що інструменти мають ряд переваг, все ж було вирішено створити базу знань, так як:

- Google Scholar немає RDF/OWL-технологій, DBPedia не є бібліотекою публікацій, а "базою" інформаційних сторінок;
- існуючі прототипи містять багато зайвих об'єктів, суб'єктів, властивостей, а це ускладнюватиме роботу тих, хто наповнюватиме базу знань;
- у розробленій базі знань реалізовано «розумний пошук» натомість як у існуючих прототипів лише за ключовими словами.

Досягнення поставленої мети реалізовується за допомогою програмного забезпечення Protege, для проектування бази знань та заповнення її даними та мови SPARQL, за допомогою якої можна робити запити в базу, тим самим здійснювати пошук по ній.

Бібліотека публікацій заповнялась публікаціями із сайту grid.kpi.ua.

На рис. 1 зображено фрагмент бази знань наукових публікацій.

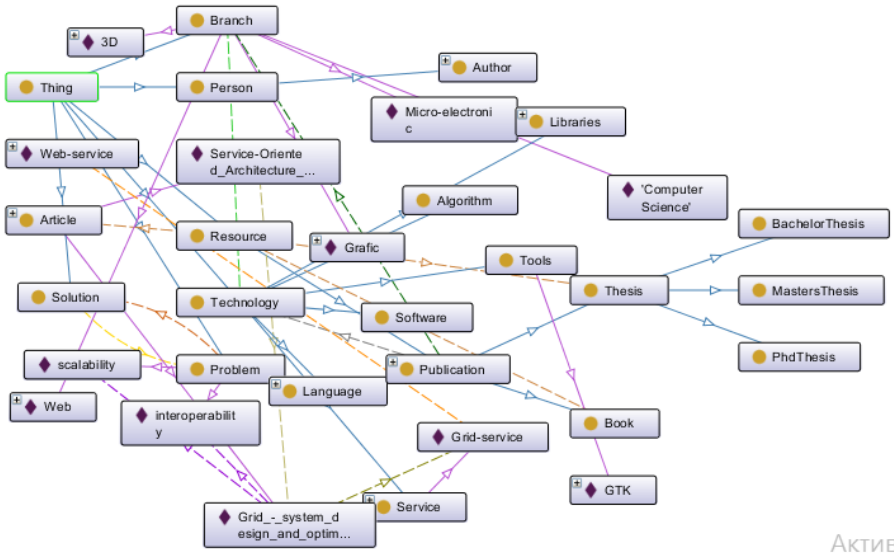

Рисунок 1 - Фрагмент бази знань(складено автором на основі [5, с. 105])

У бібліотеці публікацій є стаття: «Грід-системи з розробки та оптимізації інженерних рішень», Мета – це дослідження, предмет – системи з розробки інженерних рішень, засоби, які використовуються для проведення цього дослідження – MATLAB, Mathematica, CAD-Grid, NetALLTED, ROOT. NetALLTED – вітчизняна система для інженерного проектування. Нехай нам потрібно знайти всі статті, в яких розглядаються

вітчизняні системи для інженерного проектування. Результатом такого запиту і буде дана стаття (на рис 2).

```
SELECT ?name
WHERE
   {
       ?article lp:name ?name,
       ?article lp:subject ?subject,
       ?subject lp:use ?tool,
       ?tool lp:type ?Tool,
       ?tool lp:type ?'Native'
   }
```

Рисунок 2 - Запит, що шукає статті, в яких розглядаються вітчизняні системи для інженерного проектування (складено автором на основі [5, с. 25])

Далі у нас є стаття – «Методологія розробки онтологій», мета – показати процес створення онтології, предмет – онтологія, опис – створення онтології, що описує Грід.

Нехай є запит: «Знайти статті, де описується семантичний веб», онтологія – елемент семантичного вебу, а значить дана стаття буде результатом такого запиту. На наступному слайді показано ще один цікавий запит, який знаходить авторів, які описували базу знань Грід, так як онтологія, що описує Грід у цій статті і є базою знань, то автор, який відобразиться буде Петренко А.І., який і є автором цієї статті (на рис 3).

```
SELECT ?name
WHERE
   {
       ?article lp:name ?name,
       ?article lp:describe ?'Semantic Web'
   }
```

Рисунок 3 - Запит, що шукає статті, де описується семантичний веб (складено автором на основі [5, с. 35])

Ще одним прикладом може бути запит, який знаходить всі статті, де досліджується алгоритм Max-min. Є стаття: «Алгоритми балансування навантаження в Грід-системах», метою є дослідження, проблема –

зменшити час виконання завдань, проблемою також є забезпечити ефективність використання обчислювальних ресурсів, предметом є алгоритми балансування навантаження, а саме - алгоритм планувальника, а одним із конкретних алгоритмів є алгоритм Max-min. Таким чином результатом вище згаданого запиту є дана стаття (на рис 4).

```
SELECT ?author
WHERE
   {
       ?article lp:desription ?'Grid base of knowledge',
       ?author  lp:writeArticle ?artile
   }
```

Рисунок 4 - Запит, що шукає статті, де описується семантичний веб (складено автором на основі [5, с. 34])

І ще один приклад запиту, який виводить усіх авторів публікацій, де рішенням проблеми є веб-сервіс. Є декілька статей, де метою публікації є дослідження, предмет дослідження – програмне забезпечення грід, проблема – масштабування, проблема - інтероперабельність, рішення - є - алгоритм, рішення - є грід-сервіс, а грід-сервіс - є - веб-сервіс. Такі статті написало декілька авторів, такий запит зображено на рис 2.

```
SELECT ?autors
        WHERE {
                ?autors pb:property ?Author.
                ?problem pb:type ?Problem.
                ?service pb:type ?'Web-Service'.
                ?problem pb:resolvesBy ?service
              }
```

Рисунок 5 - Запит, що шукає авторів публікацій, де рішенням проблеми є веб-сервіс (складено автором на основі [5, с. 75])

Отже, в ході роботи було досліджено створення баз знань з використанням RDF-сховищ. Складність роботи полягала у тому, що теоретична база хоч і активно розвивається, і є перспективною, проте не надто добре вивчена.

Створення онтологій є перспективним напрямком сучасних досліджень по обробці інформації, що подається на природній мові. В рамках роботи було висвітлено поняття онтології, баз знань, наведено класифікацію онтологій, та описано різницю між ними.

Враховуючи універсальність формату RDF для опису не лише бібліографічних даних, а й смислових конструкцій змісту публікацій, було запропоновано розробити базу публікацій, що не лише включає набір ключових слів, а й описує основні положення тієї чи іншої публікації, що, в свою чергу, робить можливим постановку складніших за смислом пошукових запитів. Це може дати помітний виграш відносно існуючих пошукових механізмів, що здійснюють пошук інформації без урахування семантики слів, які входять до запиту, а також контексту, в якому вони використовуються. Так як це може дати помітний виграш відносно існуючих пошукових механізмів, що здійснюють пошук інформації без урахування семантики слів, які входять до запиту, а також контексту, в якому вони використовуються. У такого підходу є декілька недоліків: створення специфічних SPARQL-запитів та необхідність вручну “семантично анотувати” наявні у бібліотеці ресурси, і дослідження засобів автоматизації такої роботи є окремою актуальною задачею [6, с. 205]).

Було названо і проаналізовано ряд існуючих прототипів бібліотек наукових публікацій, що використовують RDF-сховища, такі як DBpedia, Google Scholar та Scirus. Не зважаючи на наявні переваги існуючих рішень, була поставлена задача розробки власної системи, яка б не включала "зайві" для бібліотеки публікацій класи, зв'язки та властивості, а також дозволяла б пошук за контекстом, а не лише ключовими словами. При цьому, при заповненні бази знань адміністратор може сконцентруватися на основних поняттях, не відволікаючись на заповнення безлічі другорядних атрибутів, які найчастіше не використовуються у пошукових запитах.

Було створено таку ієрархію класів та створено відповідні між ними зв'язки, заповнено необхідними даними (науковими публікаціями) базу знань, що все ж вдалось не використовуючи при описі самого суб'єкта деякі терміни, звертаючись в пошуку по ним отримати потрібний результат.

Надалі можна розвивати роботу, наповнити базу знань більшою кількістю публікацій, детальніше анотувати її, дослідити застосування баз знань у інших прикладних системах, дослідити та проаналізувати способи та методи автоматичного заповнення баз знань. Так як все ж недоліком є те, що заповнення баз знань потребує обізнаності оператора безпосередньо із предметною областю бази знань, що ускладнює процес, можливо це одна із причин, чому бази знань дуже рідко, на жаль, використовують саме в цілях розумного пошуку, а не як звичайну СУБД чи пошук за ключовими словами. Тому в майбутньому можна продовжувати роботу та досліджувати нові можливості онтологій, так як на даний час це відносно нові технології, але дуже перспективні.

Література:

1. Basu D. Smart Doorplate / Basu D. // Journal of Ontologies. – 2003. – №25. – С. 201-216.
2. Cardenas A.F. Data Base Communication in a heterogeneous data base management system network / Cardenas A.F. — K. : Information Systems, 2010, 253 с.
3. Garcia-Molino H. N. RDF in Semantic Web/ Garcia-Molino H. N. — K.: NGITS, 2013. — 110 с.
4. Gruber, T.R. A translation approach to portable ontology specifications / Gruber, T.R. — С. : Knowledge Acquisition, 2015. — 321 с.
5. Щербак С. В. Руководство по созданию онтологий / Щербак С. В. – К.: Техника, 2014. – 302 с.