УДК 004.5+004.822

Н.О. Біліченко, Д.В. Кисюк, Т.М. Павлик
Вінницький національний технічний унівеситет

ДОСЛІДЖЕННЯ ЗАСТОСУВАННЯ СЕМАНТИЧНОЇ МОДЕЛІ РЕСУРСІВ ДЛЯ ІНФОРМАЦІЙНОГО ПОШУКУ

Розглянуто проблему інформаційного пошуку в мережі Internet та особливості здійснення інформаційного пошуку сучасними пошуковими системами. Проведено аналіз семантичних технологій та розглянуто можливість їх застосування для розроблення ефективних систем пошуку даних у глобальній мережі.

Наведено схему роботи інформаційного пошуку на основі семантичної моделі ресурсів. Розглянуто основні семантичні структури та проаналізовано основні засоби їх побудови. Досліджено структуру семантичного стека та призначення стандартів семантичної моделі ресурсів. Проведено порівняльний аналіз існуючих семантичних движків у межах стандарту RDF.

Ключові слова: інформаційний пошук, семантична модель ресурсів, онтологія, RDF, семантичні движки.

This article deals with the problem of information retrieval on the Internet. The specific features of information retrieval implementation by modern search engines are overviewed. The analysis of semantic technologies was carried out. Possibility to use them for effective data retrieval systems development in a global network is considered. Functional scheme of semantic technologies based on semantic resources model is overviewed. The basic semantic structures are examined and the basic means of their creation are analyzed. The structure of the semantic stack and assigning of the semantic resource model standard were investigated. A comparative analysis of existing semantic engines within the RDF standard was carried out.

Key words: information search, semantic resources model, semantic ontology, RDF, semantic engines.

Вступ

Проблема пошуку інформації є однією із серйозних проблем, з якою зіткнулось сучасне “інформаційне суспільство”. Ріст обсягів даних, що зберігаються в мережі Internet, зумовлюють актуалізацію проблеми інформаційного пошуку. Сучасні засоби пошуку, опису текстів, каталогізації не можуть задовольнити потреби користувачів повною мірою. Потрібно шукати нові напрямки підвищення ефективності систем інформаційного пошуку та спрощення їх взаємодії з користувачем.

Мета роботи – провести дослідження застосування семантичної моделі ресурсів для інформаційного пошуку.

Об’єктом дослідження постає семантична модель ресурсів.

Предметом дослідженя є методи та засоби реалізації системи інформаційного пошуку на основі семантичної моделі ресурсів.

Головним завданням є аналіз семантичних технологій та їх застосування для побудови пошукових систем, здатних підвищити ефективність пошуку та обробки інформації у мережі Internet.

Постановка проблеми

У сучасних пошукових системах текстовий матеріал індексується автоматично за набором ключових слів, що входять до їх складу. Таке подання інформації як звичайного набору текстів має багато очевидних недоліків, а саме:

1. Надлишковість – пошук за ключовими словами дає занадто багато зайвих посилань.

2. Багатозначність ключових слів – багатозначні слова можуть мати кілька понять, що виражають їх різне значення, тому малоймовірно, що вони зацікавлять користувача.

3. Низька точність та надійність результатів пошуку.

У зв'язку з цим пропонується використовувати семантичну модель ресурсів, яка позбавлена вищевказаних недоліків, за рахунок використання концептуального індексування, тобто індексування за поняттями, а не за ключовими словами.

Схему роботи системи інформаційного пошуку на основі семантичної моделі ресурсів показано на рис. 1. Користувач вводить запит. На наступному етапі запит піддається лінгвістичному аналізу, перетворюється у ключові слова та відправляється пошуковій машині. Остання повертає знайдені документи, із яких після лінгвістичного розбору формуються семантичні образи документів. Отримані образи порівнюються із запитом, після чого робиться висновок про релевантність кожного із документів і результати аналізу видаються користувачеві.

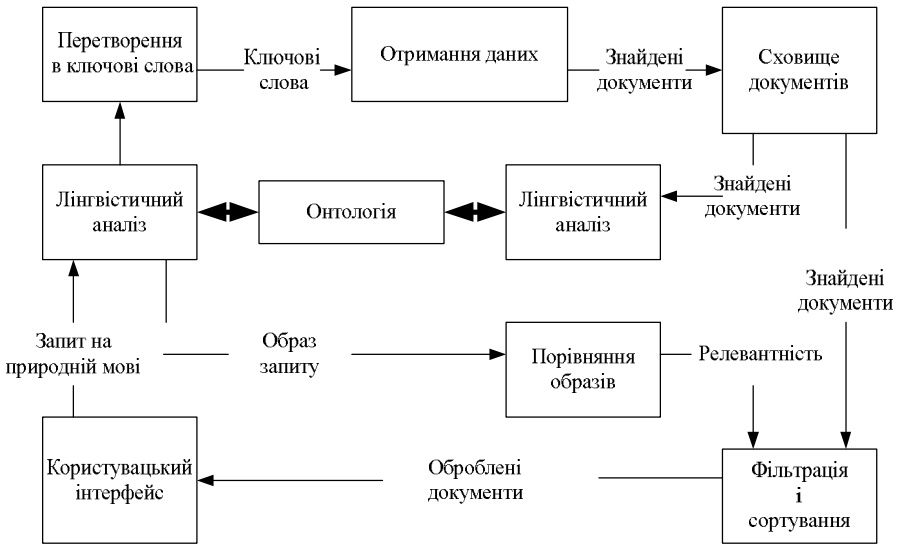

Рис. 1. Схема роботи системи інформаційного пошуку на основі семантичної моделі ресурсів

За такої технології усі багатозначні слова, віднесені до різних понять, усі синоніми, зведені до одного і того самого поняття, а зв'язки між поняттями і відповідними словами чітко описуються і можуть використовуватись під час аналізу тексту [1].

Як показують дослідження світових лідерів в прогностичних дослідженнях [Tofler, 2006; IDC, 2012, GARTNER, 2012], до 2020 року кількість інформації і потреби в ній зростатимуть експоненційно. Тому важливим є пошук нових шляхів не лише збереження і обробки даних, але й накопичення і обробки знань, що, своєю чергою, формує нову семантичну хвилю, яка, за оцінками [2], істотно змінить характер роботи з інформацією.

Ідея створення семантичних пошукових систем була висунута у 2000 році одним із основоположників WWW і нинішнім представником консорціуму W3C Тімом Бернерсом-Лі (Tim Berners-Lee) [3]. Розробленню питань, що стосуються інформаційного пошуку на основі семантичних технологій, присвячені роботи таких науковців, як: П. Ломбарді [4], Є. В. Слесарєва [5], Д. Г. Колба [6], В. Ф. Хорошевського [7], Ю. І. Шокіна [8].

Виклад основного матеріалу

Автори провели дослідження, в результаті яких було виявлено, що використання семантичних технологій для побудови пошукових систем здатне підвищити ефективність інформаційного пошуку та обробки інформації у мережі Internet.

Семантична модель ресурсів передбачає запис інформації у вигляді семантичної мережі за допомогою спеціальних семантичних структур – онтологій. І тоді, як у WWW інформація, що закладена в HTML-сторінках, добувається користувачем за допомогою браузера, у семантичній мережі програма-клієнт може добувати з Internet факти і робити з них логічні висновки.

Сучасні методи пошуку інформації в Internet ґрунтуються, як правило, на частотному і лексичному аналізі тексту. Семантична модель ресурсів натомість використовує стандарт RDF (Resource Description Framework), який описує семантичні графи, вузли і дуги яких мають уніфіковані ідентифікатори ресурсів (URI). Твердження, закодоване за допомогою RDF, в подальшому можна інтерпретувати за допомогою онтологій, створених за стандартами OWL та RDF Schema, для того, щоб отримати із них відповідні логічні висновки.

Семантичний стек містить два рівні, які мають пряме відношення до семантичних структур – це RDF-структури та онтології. Разом вони являють собою комплексну систему метаданих, яка може бути адаптована до будь-якої системи. RDF є основою будь-якої семантичної системи. Цей стандарт являє собою твердження про ресурси у вигляді, придатному для машинної обробки. Ресурсом RDF слугує будь-яка сутність (наприклад, зображення, веб-сайт). Отже, якщо потрібно описати якийсь ресурс за допомогою RDF, слід зробити ряд тверджень або фактів, як вони називаються в теорії семантичного аналізу. Із простих фактів будується структура семантичного опису ресурсу. Однак простого переліку фактів недостатньо. Необхідні правила, за якими відбуватиметься аналіз фактів і відповідне їм віднесення об’єкта до тієї чи іншої категорії. Ці правила містяться у так званих онтологіях, які описують ієрархічні відносини під різними предикатами і навіть між різними категоріями.

Для реалізації усієї повноти концепції онтології існують окремі розширення (движки), поки що не стандартизовані, однак вони розвиваються в межах одного “батьківського” стандарту RDF. До них належать:

1. R2RML Parser – модуль, призначений для використання у мові програмування Java. Дозволяє не лише “розбирати” RDF, а й генерувати нові структури у цьому форматі. Також має обмежену підтримку мова запитів у межах простих онтологій.

2. dotNetRDF – модуль, призначений для використання в мові програмування C#, та інших мовах програмування, що ґрунтуються на технології .NET. Так само, як і RDFSharp, і попередня бібліотека призначена для маніпуляцій RDF на простому рівні, а також дозволяє формулювати прості запити.

3. 4Suite – комплексний пакет для роботи з XML мовою програмування Python. Містить підтримку RDF як одного із XML-форматів. Може обробляти RDF, але не формулювати запити.

4. ActiveRDF – пакет для роботи з RDF для мови програмування Ruby. Теж дає змогу обробляти, а не формулювати запити. Є частиною веб-движка Ruby on Rails.

5. ARC RDF Store – пакет для роботи з RDF безпосередньо з веб-сторінок, написаних мовою PHP. Є частиною веб-движка.

6. Brahms – пакет для роботи з RDF з мови програмування C++. Швидкий, здатний опрацьовувати великі онтології. Являє собою велику бібліотеку класів.

Проведемо порівняльний аналіз вищевказаних движків та зведемо отримані дані у таблицю.

Порівняльна таблиця движків RDF

	Платформи	Веб-движок	Створення RDF	Обробка RDF	Запити
R2RML Parser	Усі	-	+	+	+
dotNetRDF	Windows	-	+	+	+
4Suite	Усі	+	+	+	-
ActiveRDF	Усі	+	+	+	-
ARC RDF Store	Усі	+	+	+	-
Brahms	Windows, Linux	-	+	+	+

Вибір мови специфікації онтології є ключовим моментом у проектуванні семантичної моделі. Саме тому аналіз методів та засобів RDF є вкрай важливим. Вищезазначена порівняльна таблиця движків RDF покликана допомогти розробникам під час реалізації систем інформаційного пошуку на основі семантичної моделі ресурсів.

Семантичні технології надають багато можливостей користувачеві під час його взаємодії із системою інформаційного пошуку. Адаптивні онтології, призначені для людського розуміння, покривають термінологію кінцевих користувачів, включаючи синоніми та омоніми. Це дає змогу впроваджувати прогресивні технології пошуку даних. Зрештою, за допомогою семантичних можливостей можлива і реалізація найвищої мети пошуку, коли відповідь на пошуковий запит стає не "що користувач сказав", а "що користувач мав на увазі" [9].

Висновки

Задача інформаційного пошуку в умовах галопуючого розвитку сучасних інформаційних технологій є однією із ключових у сучасній теорії інформації та має фундаментальне значення у процесах обробки даних. Оскільки традиційні підходи інформаційного пошуку у мережі Internet стають все менш ефективними через зростання обсягу інформаційних ресурсів, то постає необхідність у пошуку та розробленні альтернативних методів і засобів, які б уможливили швидше та ефективніше розв'язувати поставлені задачі пошуку інформації. Результати проведеного дослідження семантичної моделі ресурсів для інформаційного пошуку підтверджують важливість розроблення ефективних пошукових систем, орієнтованих на спрощення задач пошуку та обробки даних.

1. Козлов Д. Д. Информационно-поисковые системы в Internet: текущее состояние и пути развития. – М., 2000. 2. Mills D. Semantic Wave 2006. Executive Guide to Billion Dollar Markets. A Project10X Special Report. January 2006. 3. Berners-Lee, T. The Semantic Web and Research Challenges – http://www.w3.org/2003/Talks/01-sweb-tbl/slide1-0.html. 4. Paolo Lombardi. In the heart of Semantic Technology. – Italy, 2009. – http://www.elda.org/medar-conference/pdf/8.pdf. 5. Слесарев Е.В. Преимущества семантических технологий: практический аспект. – Саранск, 2012. – http://fetmag.mrsu.ru/2012-1/pdf/Slesarev.pdf. 6. Колб Д.Г. Web-ориентированная реализация семантических моделей интеллектуальных систем. – Минск, 2012. – http://conf.ostis.net/images/d/dc/Колб_Д.Г.2012ст-WebОРСМИС.pdf. 7. Хорошевский В.Ф. Семантические технологии: ожидания и тренды. – М., 2012. – http://www.hse.ru/pubs/lib/data/access/ticket/138329255690c430c68f3472c45f1aad5b737abd1c/Хорошевский (OSTIS-2012).pdf. 8. Шокин Ю.И. Проблемы поиска информации. – М.: Изд. дом. "Наука Н", 2010. 9. Ландэ Д. Семантический веб: от идеи к технологии. – http://poiskbook.kiev.ua/sw.html.