УДК 519.688

К.т.н., доцент Петрашенко А.В., магістрант Бояркін К.К.

**Національний технічний університет України
«Київський політехнічний інститут»**

АНАЛІЗ ВЕБ-ДОКУМЕНТІВ ІЗ ВИКОРИСТАННЯМ RDF-МОДЕЛІ

Abstract

Andrij V. Petrashenko, assoc. prof., PhD; Kyrylo Boiarkin, student

Analysis of web documents using RDF model

This paper concerns a method for automatic generation of descriptive information in the RDF for the web sites. The method is based on an analysis of a web document and the corresponding database, the construction of keywords graphs and finding of new connections by merging graphs. This will automatically convert web site into publish extensive information about content of their pages that will lead to more efficient computer processing of the site's content.

Вступ

За останні десятиліття глобальна мережа Internet розвивається настільки швидко, що існуючих засобів пошуку та систематизації інформації стає недостатньо і тому виникає необхідність у розробці нових алгоритмів структуризації, пошуку та обробки даних.

Один із підходів до систематизації ресурсів у мережі Internet пов'язують із концепцією семантичного Веб, що дозволить більш якісно оброблювати гіпертекстову інформацію за допомогою спеціальної мови розмітки документів *RDF*.

Постановка задачі

Виходячи з того, що семантичний Веб на сьогоднішній день недостатньо опрацьований і досить мала частка сайтів надає описовий документ *RDF* для його вмісту, актуальною є задача автоматичного аналізу та побудови *RDF*-моделі подання даних про документ.

Термінологія

Семантичний Веб – спосіб систематизації інформації в мережі Internet, завдяки якому дані можуть бути представленими у формалізованому вигляді, що дає можливість людям і комп'ютерам працювати з більш високим ступенем взаєморозуміння і узгодженості [1].

RDF – розроблена консорціумом *W3C* технологія семантичного Веб, яка включає в себе середовище опису ресурсів (англ. *Resource Description Framework, RDF*), визначає загальну архітектуру метаданих і призначена для забезпечення сумісності метаданих за допомогою спільної семантики, структури та синтаксису [2].

Кодування метаданих на мові RDF

RDF є універсальним методом поділу знань на маленькі частини відповідно до певних правил, що враховують семантику (суть) цих частин. Кожен *RDF* документ складається з трійок даних: "об'єкт - предикат - суб'єкт". Маючи такий описовий документ, Веб-сайт зможе описати інформацію, яка знаходиться на його сторінках, що дозволить стороннім програмним комплексам більш інтелектуально оброблювати ці дані, а також підвищить релевантність результатів пошуку [3].

На даний момент існують алгоритми побудови цих документів на основі розмітки мовою *XHTML*, але цей метод не дає можливості описати семантику інформації, що знаходиться на відповідних сторінках сайту. Іншим поширеним методом є побудова документу програмістом під час проектування та розробки Веб-сайту – це дає змогу отримати відносно якісний *RDF*-документ з описом семантики інформації. Недоліком цього методу є те, що не кожен програміст на практиці будує такі документи. Крім того, на даний час існує дуже велика кількість функціонуючих Веб-сайтів, для яких побудова RDF у ручному режимі не є доцільною.

Алгоритм автоматичної побудови RDF документа

RDF-документ є графом, який може мати зв'язки з іншими графами, тобто з іншими *RDF* документами. Будь-який текст, який міститься у Веб-сторінці, можна представити у вигляді графу, де вершинами будуть ключові слова, які часто зустрічаються на сторінці, а дугами – логічні зв'язки між словами. Основною проблемою є відсутність засобів автоматичного знаходження таких логічних зв'язків між словами в документі.

З іншого боку існує інше джерело зв'язків між фрагментами тексту – інформаційна база даних. Програміст вже при проектуванні бази даних задає логічні зв'язки між таблицями (сутностями), тим самим і між інформацією, яка там зберігається. При цьому, базу даних теж можна представити графом, де роль вершин відіграють рядки або поля в таблиці, а дугами будуть зв’язки між таблицями. Таким чином, маючи граф ключових слів на Веб-сторінці та граф бази даних, можна побудувати зв'язки між ключовими словами, що дозволить створити *RDF*-документ, який описує Веб-сторінку.

Кожен документ можна представити як набір ключових слів та зобразити їх на графі без зв'язків. Вибираючи в тексті послідовно ключові слова, з них можна побудувати граф – G_{keywords}. При цьому можна враховувати відстані між вершинами – чим ближче вершини, тим ближче ключові слова знаходяться в тексті, і тим більш імовірним буде зв'язок між ними.

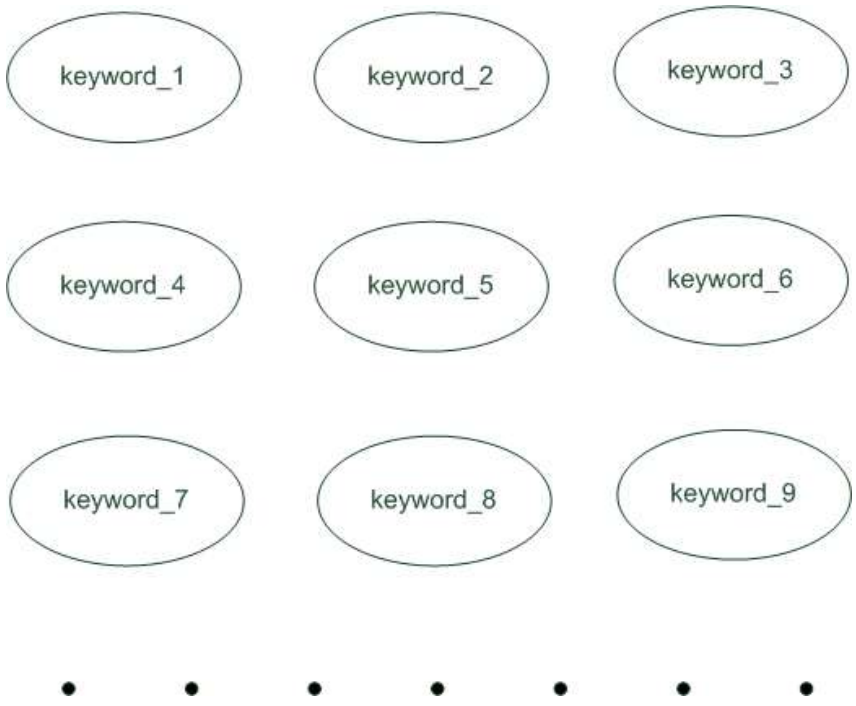

Рис. 1. Граф ключових слів, знайдених у тексті сторінок Веб-сайту

Веб-сайти зберігають свою інформацію в базі даних у вигляді таблиць, записів та полів. Записи в таблицях мають між собою зв'язки, закладені розробником бази даних. На основі структури бази даних можна побудувати інформаційний граф G_{db}.

Далі, об'єднуючи графи ключових слів документа G_{keywords} і граф логічних зв’язків даних G_{db}, можна отримати описову інформацію про елементи в документі:

$$G_{\text{main}} = G_{\text{keywords}} \cup G_{\text{db}}$$

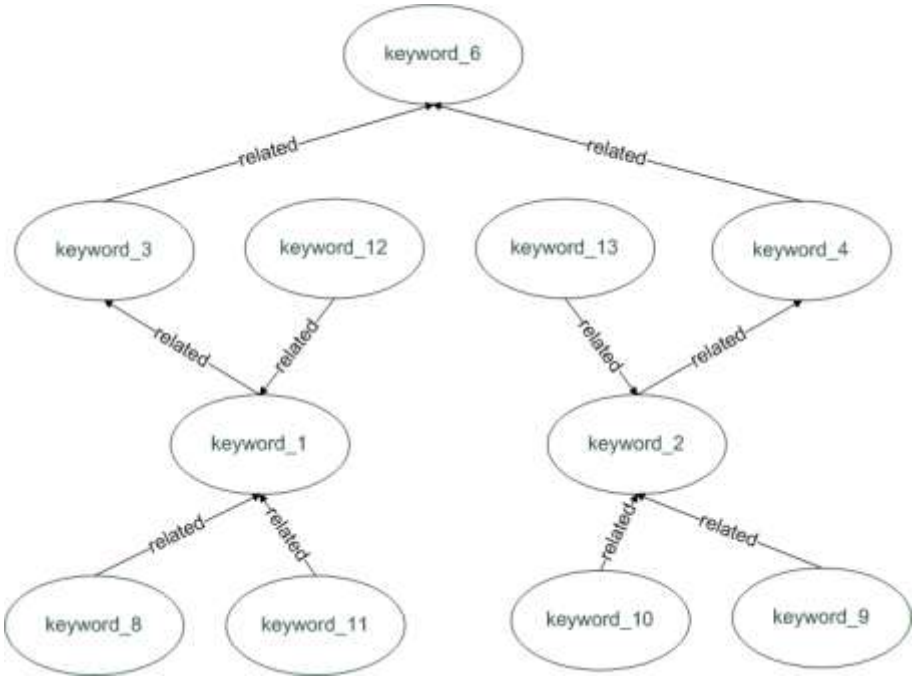

Рис. 2. Граф опису ключових слів Веб-сайту

Маючи граф, що описує семантику ключових слів G_{main} можна побудувати *RDF*-документ, що містить опис Веб-сайту. Фрагмент такого документа наведено нижче:

```
<html xmlns:foaf="http://xmlns.com/foaf/0.1/">
    <div xmlns:table1ns="http:// table1ns.org/keyword_1/" about="table_1:keyword_1">
        <span property="foaf:table_1_field_1">keyword_1</span>
        <span property="table_2:table_2_field_1">keyword_3</span>
    </div>
</html>
```

При цьому, атрибут *about="table_1:keyword_1"* визначає, що інформація, яка знаходиться всередині цього елемента, відноситься до запису *keyword_1* таблиці *table_1*. А запис *xmlns:table1ns="http://table1ns.org/keyword_1/"* – це Веб-адреса, де знаходиться загальнодоступний список ідентифікаторів, до числа яких входять записи, що зберігаються у таблиці *table_1*.

Висновки

У ході проведеної роботи проаналізовано можливість автоматичного подання інформації на Веб-сайті у вигляді *RDF*-документа. Запропонований метод аналізу ключових слів на сторінці та взаємозв'язків в базі даних.

Як результат цього аналізу виникає можливість визначати зв'язки ключових слів на сторінці, знаходити логічний зміст тексту і описувати елементи сторінок Веб-сайту.

Крім того, важливим аспектом даної роботи є досягнення можливості опису Веб-документу засобами *RDF*-формату в автоматичному режимі, що підвищує ефективність застосування засобів семантичного Веб.

Література

1. *Dieter Fensel, Jim Hendler, Henry Lieberman, and Wolfgang Wahlster.* Introduction to Spinning the Semantic Web. - Cambridge: MIT Press 2003, 1 - 25.
2. *Shelley Powers*. Practical RDF. - O'Reilly, 2003, 15-21.
3. *John Hebeler, Matthew Fisher, Ryan Blace, Andrew Perez-Lopez*. Semantic Web Programming. - Wiley Publishing, Inc, 2009, 137 - 151.
4. *Melike Şah, Wendy Hall and David C De Roure*. SemWeB Semantic Web Browser – Improving. Browsing Experience with Semantic and Personalized. Information and Hyperlinks. - School of Electronics and Computer Science, University of Southampton, 2009.