УДК 681.3.06

К.т.н., доцент Білостоцький А.І., студент Курутін М.С.

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

МЕТОД СЕМАНТИЧНОГО ПОШУКУ НА ОСНОВІ ГЕНЕРАЦІЇ ОНТОЛОГІЙ ВЕБ-ДОКУМЕНТІВ

Abstract

Anatoliy Belostotskiy, assoc. prof., PhD; Mykola Kurutin, student
Semantic search method based on web-documents ontology generation

This paper offers a new method of semantic search based on generations of ontology from existing context of web-documents that were browsed before. In order to complete ontology generation Latent Semantic Indexing method is used. This document also provides a scratch for building an internet-based semantic search system.

Вступ

Звичайно, пошукові машини Інтернет допомагають користувачеві прискорити процес інформаційного пошуку, але часто відсутність контексту пошукових слів перешкоджає ефективності послуг інформаційного пошуку. Ці проблеми вирішуються шляхом впровадження технології XML, яка була розроблена з метою формування опису структури та семантики даних.

Іншою проблемою є знання про інтереси користувача або контекст в процесі інформаційного пошуку. Пошукові машини віднаходять та класифікують знайдену інформацію на основі ключових слів та їх характеристик. Отримується велика кількість сторінок, що містять ключові слова, проте сторінки можуть не мати необхідної інформації для користувача. Пошукові слова характеризуються множинами значень (проблема полісемії) [1]. Контекст, в якому ці слова з’являються, допоможе відрізнити найбільш відповідне значення запиту.

Декілька років тому американська науково-дослідницька корпорація Telcordia (колишня Bell Labs) розпочала дослідження можливостей покращення веб-пошуку шляхом використання відомостей про потреби користувача [2], [3], підтримки моделі його інтересів на стороні клієнта та використання їх для фільтрації найбільш відповідних запиту результатів. Дана стаття продовжує та розвиває цей напрямок досліджень. Зокрема автори пропонують включити семантичне доповнення, що відображає

інтереси користувача, безпосередньо у пошуковий запит. Для реалізації цієї мети розроблена модель генерації доповнення з документів користувача.

Постановка задачі

Метою даної статті є запропонувати новий підхід для покращення існуючих інформаційних пошукових систем шляхом додання семантичних розширень, які посилюють якість послуг інформаційного пошуку.

Об'єктом дослідження є множина веб-документів доступна у глобальній мережі Інтернет та представлена у текстових форматах HTML, XHTML та XML.

Предметом дослідження є методи виділення з текстових документів семантичних концептів конкретної тематичної онтології, а також методи обробки цих концептів.

Вибір методів та модель розв'язку

У процесі досягнення цієї цілі будемо використовувати сучасні статистичні методи інформаційного пошуку, зокрема метод латентного семантичного індексування (Latent Semantic Indexing - LSI), за якими робиться спроба зрозуміти статистичні посилання термінів шляхом заміни простору термів документа на значно менших розмірів простір концептів. В LSI це виконується використанням методу матричної декомпозиції – Singular Value Decomposition (SVD). Ефективність SVD у порівнянні з іншими методами описана в [4].

Для побудови конкретної онтології використовується послідовність наступних процедур: читання збірки відповідних текстових документів та вилучення онтологічної інформації статистичними методами шляхом попередньої обробки текстів документів (Pre-processing), нормалізації текстів (Normalization), формування семантичних концептів, що відносяться до значимих термінів, з використанням LSI та SVD. Модель семантичної пошукової системи, що працює згідно такого підходу зображена на рис. 1.

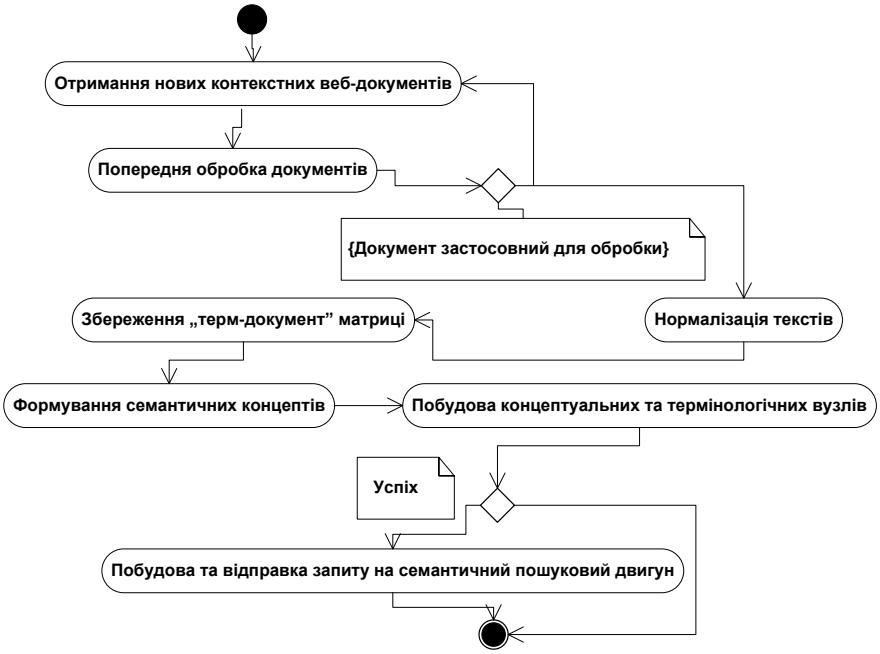

Рис. 1. Модель семантичної пошукової системи. Діаграма діяльності в нотації UML

Попередня обробка є процесом, в якому здійснюється здобуття значимих термінів та підраховуються їх частоти під час читання чи завантаження текстового файлу. Над текстовим документом виконуються кілька процедур представлення тексту в необхідному форматі для впевненості у тому, що підрахована статистика є значимою. Ці процедури стосуються визначення загальних основ слів та відсічення відмічених, що семантично малозначущі.

Нормалізація - це процес, за якого рахується нормалізована вага кожного слова, яке було отримано в результаті попередньої обробки. В попередній обробці документ аналізується в аспекті корелятивних слів та матриці частотності, відомої як матриця „терм-документ", що створюється в результаті проведення аналізу.

Першим кроком є визначення числа частотності кожного терму в документі. Потім рахується вага кожного терма нормалізована відповідно до множини документів за наступною формулою:

$$NW(i,k) = \frac{W(i,k)}{\sqrt{\sum_{j=1}^{n_k} W^2(i,k)}}$$

де *W(i, k)* - вага *i*-го терма в *k*-му документі; n_k - загальна кількість термів в документі; $f_r(i, k)$ – частота терма *i* в документі *k*.

З використанням методу латентного семантичного індексування (Latent Semantic Indexing - LSI) створена матриця розкладається на три

матриці: термів *(U)*, одинична діагональна *(S)* та документів *(V)*. Після мінімізації об'єму матриць матриця термів розкладається на вектори термінів, визначених як концепти, що формуються як група відповідних термінів.

Побудова онтології документа є, по суті, побудовою концептуальних та термінологічних вузлів з матриці термів *(U)* і документів *(V)*. Онтологія представляється у вигляді графа. Сформований граф використовується, щоб показати зв'язки між різними термінами та концептами. Концептуальні вузли поєднані з термінологічними, які непрямо пов'язані з іншими концептуальними. Термінологічні вузли зв'язані з іншими термінологічними вузлами тільки шляхом зв'язування з вузлом загального концепту.

Концептуальний вузол представляє концепт і містить інформацію про його назву, терми, що відносяться до нього, та їх ваги в концепті.

Висновки

У статті запропоновані підхід створення онтології з текстових документів з використанням методу LSI, який будує предметно-орієнтовану онтологію із збірки текстових документів. Метод, у цій конструкції є статистичним. Він застосовує добре відому матричну декомпозицію та надає результати, дійсність яких підтверджується теоретично. Система забезпечує швидке формування предметно-орієнтованої онтології для використання її в якості запиту в інформаційно-пошукових системах.

Один з можливих напрямків покращення якості пошуку, полягає у визначенні семантики зв'язків концептів та в використанні зв'язків між концептами в якості запиту.

Список літератури

1. *Miller G.A., Beckwith R., Fellbaum Ch.,Gross D., Miller K.* Introduction to Word-Net: An On-line Lexical Database // *International Journal of Lexicography*. –2006. –17(2). – P. 72-95.
2. *Bassu D., Behrens C.* Applied Research Distributed LSI // *International Workshop on Research Issues in Data Engineering*. – 2001. – P. 52-58
3. *Chen C., Stoffel N., Post M.* Telcordia LSI Engine: Implementation and Scalability Issues // *Telcordia Diagest*. – 2004. – 13(1). - P. 18-47
4. *Berry M.W., Dumais S.T., O'Brein G.W.* Using linear algebra intelligent information retrieval // *SIAM Review*. –1995. –37(4). – P. 573-595.