УДК 681.3.04

д.т.н., професор Зайцев В.Г., магістрант Лесько С.В.

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

ОСОБЛИВОСТІ РЕАЛІЗАЦІЇ СЕМАНТИЧНОГО ПОШУКУ

Abstract

Volodymyr G. Zaitsev, prof., DcS; Sergiy Lesko, student
Features of implementation semantic search engine.

The essay is dedicated to comparative analyze of existing search methods and usage of semantic search engine. In this work semantic method analyze is proposed and described it's comparing with relevant method search.

Вступ

На даний момент в переважній більшості систем пошуку використовується релевантна система оцінки відповідності досліджуваного документу пошуковому запиту. Релевантна модель практично не справляється з вирішенням поставлених задач розпізнавання і пошуку омонімів (граматичних і, особливо, лексичних), синонімів та багатозначних слів. Це зумовлено тим, що в основу релевантних моделей пошуку закладений лінгвістичний підхід і ряд оціночних синтетичних критеріїв (таких як положення слів на сторінці), а перераховані вище мовні одиниці не можуть бути розпізнані без осмислення значення пошукового запиту. Семантичні пошукові системи намагаються внести такий смисл в результати пошуку, поданого в контекстному форматі.

Постановка задачі

Задача полягає в дослідженні відмінності семантичного пошуку від існуючих релевантних моделей, обґрунтуванні доцільності подальшого розвитку даної моделі, а також знаходженні способу оцінки продуктивності алгоритму семантичного пошуку, представленого в цій статті.

Термінологія

Релевантний (повнотекстовий) пошук – пошук по всьому змісту документа. Для прискорення пошуку, зазвичай, використовуються попередньо побудовані індекси, в яких зазначається інформація щодо кількості та місця конкретного слова в тексті [1, 2].

Семантичний пошук – це засіб покращення пошуку завдяки розумінню намірів користувача і контексту пошуку, що дозволяє створювати пошукові запити близькі до питань, які задаються природною мовою [3].

Порівняльний аналіз існуючих моделей пошуку

До найпоширеніших методів пошуку відносять наступні:

Булевий пошук [4]– це комбінація елементів, що дозволяють включати і виключати із пошукових результатів документи, що містять певні слова. Це досягається за допомогою булевих операторів and, not, or, near. Він являє з себе найпростішу **з** пошукових програм порівняння. В найпоширеніших пошукових сайтах такий тип пошуку реалізований уже давно. Пошук із застосуванням **wildcard символів** [5], властиві багатьом пошуковим системам. Найчастіше wildcard символи використовуються у вигляді символа астериска (*) або знака питання для заміни букв чи послідовностей букв при написанні. Деякі системи підтримують **пошук з відстанню** [6], де слова можуть знаходитися на певній відстані одине ід другого. Для здійснення такого пошуку використовують значок тильди (~). Неточний пошук - це особливий вид пошуку в процесі якого визначаються сторінки, які можуть бути релевантні аргументу пошуку, навіть якщо аргумент не точно відповідає бажаній інформації. Неточний пошук відбувається за допомогою програми, яка демонструє список результатів, побудованих на основі деякої схожості з заданим варіантом. Найбільш точні та релевантні співпадіння можуть знаходитися на початку всього списку результатів пошуку.

Пошук по контексту [7]**.** Пошук по контексту є електронною версією визначення слова в залежності від оточуючих його лексем (контексту). Даний вид пошуку має часткову схожість з неточним пошуком з відмінністю в тому, що пошук по контексту передбачає оцінку змісту всього тексту, а не окремих слів.

В даній роботі пропонується використання семантичної пошукової моделі. Принципова різниця між релевантною і **семантичною пошуковою** моделями полягає у тому, що при релевантному пошуку текст розглядається з точки зору форми, а при семантичному пошуку - з точки

зору змісту. Це означає, що в релевантній моделі маємо лише деякий екстракт з документу, що зберігається в базі даних, разом з посиланням на даний документ. При змістовному (семантичному) пошуку оперуємо всім змістом документа, для визначення його змісту і після цього формуємо представлення про його релевантність.

Інформаційна база зберігається в **тезаурусах** - це різновид словників загальної чи спеціальної лексики у яких вказані семантичні відношення (наприклад, синоніми, антоніми, пароніми, гіпоніми, гіпероніми) між лексичними одиницями. Таким чином, тезауруси, особливо в електронному форматі, являються один із дієвих інструментів для опису окремих предметних областей. На відміну від тлумачного словника, тезаурус дозволяє виявити смисл не тільки за допомогою визначення, але і за допомогою співвідношення слова з іншими поняттями і їх групами, завдяки чому він може бути використаний для наповнення баз знань систем інтелектуального пошуку.

Основою для реалізації семантичного пошуку є побудова орієнтованого графу, вершини якого представляють собою поняття, а ребра – визначають семантичне відношення між даними поняттями. Структурно граф може бути як деревоподібним так і містити в собі циклічні структури.

Незважаючи на те, що термінологія і структура різних семантичних методів може відрізнятися, існують деякі особливості, котрі властиві практично всім семантичним мережам (графам):

1. Вузли семантичних мереж являють собою концепти (праобрази) предметів, подій, станів;

2. Різні вузли одного концепта відносяться до різних значень, якщо не помічено, що вони відносяться до одного концепта;

3. Дуги семантичних мереж створюють відношення між вузлами-концептами (тип дуг вказує на тип відношення);

4. Деякі відношення між концептами вказують на лінгвістичні зв’язки (синоніми, антоніми та ін.), а деякі вказують логічні, просторові, часові відношення.

5. Концепти організовані по рівням залежності від ступеня узагальнення, так як, наприклад, сутність, жива істота, тварина, хижак.

Попередні розрахунки [8] показують, що при використанні пошукових запитів розміром 3-5 слів продуктивність алгоритму майже така сама як і в релевантних моделях (приблизно 70-72%), але якщо не обмежувати кількість слів у пошуку, то продуктивність пошуку семантичного алгоритму значно збільшується (82-85%).

Завдяки тому, що при аналізі використовуються додаткові метадані концептів, повнота пошуку також в середньому більша [8] не залежно від розміру пошукового запиту.

Висновок

Основною перевагою запропонованої семантичної пошукової системи є можливість задавати запити на природній людські мові. Немає необхідності виділяти ключові слова і формулювати запити в зрозумілому для машини вигляді. Робота з семантичним пошуком дуже проста. Користувач задає питання на звичайній мові, а у відповідь отримує відповіді, які позбавлені інформаційного непотребу. Крім того сама система зберігатиме історію запитів конкретного користувача, що дозволить виводити найбільш очікувані результати в кожному конкретному випадку.

Література

1. *М. С. Агеев, Б. В. Добров*. Метод эффективного расчёт матриц ближайших соседей для полнотекстовых документо // Вестник Санкт-Петербургского университета. -2011. Вып. 3. –С.72-84.
2. *Нгуен Ба Нгок, А.Ф. Тузовский.* Обзор подходов семантического поиска // Управление, вычислительная техника и информатика.-2010. №2. –С. 234-237.
3. *Dwinger, Philip*. Introduction to Boolean algebras. Würzburg: Physica Verlag. – Physica-Verlag. – 2006. –P. 311-313.
4. *Alexander Rubin*. Search with distance. MySQL AB. -2008. – С.119-122.
5. *Ahmet Soylu, Patrick De Causmaecker and Piet Desmet.* Context and Adaptivity in Pervasive Computing Environments: Links with Software Engineering and Ontological Engineering // Journal of Software. Vol.4. № 9. – 2009. 992-1013.
6. *Воробьев В.И., Перминов С.В.* Семантический поиск в интернете с использованием метаданных // Санкт-Петербургский інститут Информатики и Автоматизации РАН. -2008. – С.11