УДК 004.424.4, 004.738.52

В.О. Мокрова, А.Й. Савицький

ПОСТАНОВКА ЗАДАЧІ ПОБУДОВИ ПОШУКОВОЇ СИСТЕМИ З СЕМАНТИЧНИМ АНАЛІЗАТОРОМ НА ОСНОВІ НЕЙРОННИХ СІТОК

Вступ

Завдяки глобалізації та розширенню системи Internet, інформаційних ресурсів, можливих для використання при вирішенні різнорідних задач управління підприємством чи організацією, стає з кожним днем все більше і більше. Для вирішення конкретної задачі користувач шукає необхідну йому інформацію. Тут ключове слово "необхідну". Тому, по-перше, слід формалізувати поняття "необхідної" інформації для користувача, а по-друге – запропонувати структуру відповідної пошукової системи (ПС).

Перш за все, при формалізації поняття "необхідної" інформації користувач повинен сам розуміти, що він намагається знайти. По-друге, необхідно створити структурований природно-мовний запит для отримання належної інформації. Структура та параметри запиту повинні відповідати визначеним критеріям. В якості критеріїв, як правило, обирають релевантність та пертинентність отриманої від пошукової системи інформації (документів). При формуванні природно-мовного запиту користувач повинен розуміти різницю між основним (ключовим) словом або словосполученням та описовим чи допоміжним. Користувачу необхідно ознайомитись з можливими засобами та параметрами пошуку, що надає пошукова система, та ефективно їх використати, а пошукова система повинна адекватно інтерпретувати отриманий запит, реалізувати семантичний пошук та видати бажану інформацію. Тобто, знайдена інформація повинна відповідати потребам користувача, бути точною та повною.

Постановка задачі

На основі аналізу можливостей сучасних найпоширеніших пошукових систем визначити складові природно-мовного запиту, які слід враховувати при семантичному пошуку, а також запропонувати структуру пошукової системи з семантичним аналізатором.

Рішення

У таблиці 1 наведені результати аналізу прийомів, які використовують сучасні пошукові системи, та підходів, які повинні бути реалізовані в пошукових системах з семантичним пошуком.

В таблиці 2 наведені результати аналіз поширених пошукових систем у вигляді коротких описів їх функціональних можливостей та недосконалостей з точки зору отримання користувачем релевантної інформації.

Таблиця 1

	Сучасні ПС	**ПС з семантичним пошуком**
1	Пошук наявних слів запиту у тексті, їх взаємне розташування, що використовується лише для тих слів, які користувач ввів у пошуковий рядок.	Необхідне використання методів індексування інформації, орієнтованих на семантичний пошук.
2	Орієнтація на кількість посилань з та на даний сайт, індекси цитування використовуються лише для найпопулярніших тем, але дана характеристика не є ознакою змістовності даних.	Необхідне використання антиспамових алгоритмів пошуку інформації.
3	Використання форматування сторінки (кількості абзаців, кількості та види заголовків та ін.) як значущий критерій відбору інформації.	Необхідне включення тематичного фільтру запитів користувача (на основі множини термінів онтології предметної області) для подальшого пошуку в потрібних предметних областях.
4		Після синтаксичного та морфологічного аналізу необхідне проведення реляційно-ситуаційного аналізу для виявлення значень слів та словосполучень (синтаксем) та семантичних зв’язків між ними.
5		Для зменшення зашумленості результату пошуку необхідно використовувати метадані знайденого пошуковою системою ресурсу. До метаданих можна віднести як додаткові терміни від користувача з відповідною вагою, так і метаописи інформації розробниками сайтів.
6		На кожному етапі пошукова система повинна використовувати спеціальні словники: перелік іменників, зв’язки словосполучень за змістом, метадані категорій тощо.

 ISSN 1562-9945

Таблиця 2

ПС	Короткий опис
Google	Найпоширеніша система. Особливості пошуку: за запитом "site:<адреса сайту>", відповідь буде шукатися саме за даною адресою; за даною адресою; якщо фраза задана у лапках, то вона повинна зустрітися на сторінці повністю; використовуються багато інших службових символів. Функція "cache" є, з однієї сторони, зручною, оскільки з певної причини користувач не має змоги звернутися до серверу сторінки або до її самої, з іншої сторони, користувач одразу розуміє, що пошук по тільки що відображеній інформації на інших Internet-сторінках в даному випадку вестися не буде. Система дозволяє шукати файли певних форматів, але пошук все одно не відповідає потребам користувача у точності відповіді.
Yahoo	Система реалізує пошук сторінок лише по їх назвам та опису – не має повнотекстового індексу. Також існує проблема з сторінками, що містять текст на мові з слов'янської групи мов, оскільки в Yahoo Inc. їх не перевіряють.
Rambler	Функція «Перев'язка» є кроком вперед з точки зору семантики, оскільки дозволяє шукати сторінки не лише за словами запиту, але й за їх синонімами. Але пошук все одно не гарантує чітку необхідну відповідь та не вирішена задача «відсіювання» реклами.
Апорт	Система має механізм перекладу запитів на інші мови та має механізм реконструкції всіх проіндексованих сторінок зі своєї бази, через що користувач може потрапити на вже неіснуючу інформацію не усвідомлюючи цього.
Yandex	В системі використані деякі, відмінні від стандартних, прийоми: символ & ставиться між тими словами у запиті, які повинні бути у відповіді в одному реченні; символ «–» використовується перед словом, який потрібно вилучити з відповіді на запит; знак «!» - для пошуку слова у вказаній у запиті формі. Взагалі, ця система характеризується як одна із кращих щодо якості пошуку та технічною ефективністю порівняно з іншими існуючими на даний момент системами. Але аналіз, що проводиться цією системою, все ж таки не вирішує задачу отримання релевантної відповіді на запит.

Проблема відповідності отриманих результатів запиту є головною проблемою серед сучасних пошукових систем. Звичайно, системи використовують спеціальний синтаксис, логічні оператори, розбиття інформації на декілька категорій, використання зарезервованих слів, але цього недостатньо для належного опрацювання природно-мовного запиту.

Одним з кроків до вирішення цієї проблеми - створення вузько направлених ПС, які орієнтовані на пошук конкретного сегменту інформації. Наприклад: FindSounds.com - спеціалізується у сфері пошуку різноманітних звуків та музики у різних форматах; AllDll.net – представляє собою пошукову систему по найбільш популярним бібліотекам dll.

Але навіть, якщо запит заданий по всім можливим правилам відповідної пошукової системи, остання не гарантує однозначно-необхідну користувачеві відповідь. Звичайно, це залежить від популярності обраної користувачем теми у середовищі, де відбувається пошук, а також від правильності написання запиту самим користувачем. Але, все ж таки, пошукові системи не реалізують потреби користувачів і пертинентність залишається набагато меншою за очікувану. Релевантність у сучасних пошукових систем присутня, але лише формально. Наприклад, відбувається знаходження тих сторінок сайтів, на яких кількість слів, що записані у запиті, більша, так як вони більш популярні у середовищі пошуку та, можливо, за іншими критеріями, які дозволяють ранжувати сторінки з точки зору синтаксису. Тому релевантність за змістом відповідає потребам користувача лише при популярних запитах простої форми.

Можна приблизно представити формулу релевантності (1), яку використовують сучасні пошукові системи:

$$\text{Релевантність} = TF * IDF, \tag{1}$$

де $TF = n_i(n_i + k_1 + k_2 * DocLength)$ – пряма частота входження слова в документ,де

n_i – кількість згадувань слова у документі,

k_1, k_2 – постійні числові коефіцієнти,

$DocLength$ – довжина документу в словах;

$IDF = \log(|D| / |(d_i \supset t_i)|)$ – обернена частота документу відносно запиту, де

$|D|$ – загальна кількість документів в базі ПС,

$$|(d_i \supset t_i|$$

– кількість документів в базі, які містять дане слово t_i.

Таким чином, бачимо, що сучасні ПС здійснюють пошук синтаксично, а не семантично.

Для вирішення даної проблеми, для якісного аналізу запиту і надалі знаходження й представлення користувачу релевантної інформації, пропонується включення до програмного забезпечення пошукової системи семантичного аналізатору, який би враховував всі підтипи семантичного розбору мовних конструкцій.

При реалізації семантичного аналізатора повинні враховуватись особливості природної мови. Для цього необхідно передбачити наступне:

1. Морфологія. Пошукова система повинна мати змогу виділяти у відповідних словах відмінки, дієвідмінювання, визначати рід (чоловічий, жіночий, однину чи множину та ін. Деякі сучасні пошукові системи вже намагаються реалізувати цю функцію, але, використовуючи лише синтаксичний аналіз, вони не дають бажані результати.

2. Омоніми. Пошукова система повинна не лише "розуміти", що дане слово має декілька різнобічних значень, а й, аналізуючи інші слова у запиті, визначати, яке саме значення використовується.

3. Синоніми. Пошукова система повинна не лише "зрозуміти" значення заданого, а й визначити, які з його синонімів підходять до даного контексту, та у такому ж обсязі виконувати пошук і по ним. Сучасні пошукові системи намагаються використовувати синоніми, але лише тоді, коли по ключовим словам запиту майже нічого не знайдено. Це не вірний підхід, оскільки часто буває, що сторінка зі словами-синонімами більше підходить за змістом, а ніж ті, що були на перших позиціях у результатах за синтаксичним пошуком.

4.Онтологія. Дозволяє формалізувати знання як користувача, так і пошукової машини в конкретних областях знань. Пошукова система повинна "розуміти" ключові слова або терміни, які задав користувач при пошуку в конкретній предметній області.

5. Семантичний аналіз як такий. Тут мається на увазі обробка запиту не по словам, а за змістом. Наприклад, якщо шукається назва певного твору, то користувач повинен або вводити "назва твору", при чому пошукова система видасть в результат ті сторінки, де є саме словосполучення "назва твору" присутнє, і не видасть ті, де це слово відсутнє, хоча сама назва твору присутня. Або ж користувач не вводить це слово, а лише саму назву твору, яку пошукова система звичайно ж не ідентифікує так, як потрібно, а проводить синтаксичний пошук по словам запиту.

Таким чином, пошукова система повинна використовувати метадані (такі як "назва твору"), якими користувач міг би уточнювати свій запит, надавати системі додаткову інформацію, яка б дозволяла проводити пошук саме за значенням запиту.

В якості таких допоміжних метаданих пропонується використовувати поняття "Асоціативність", яке включатиме в себе і поняття "ієрархічність", і "категорійність". Їх використання носить лише допоміжний характер. Користувач, в свою чергу, повинен усвідомлювати різний рівень значення слів та словосполучень, їх приналежність до тої чи іншої області знань. Користувач міг би вибирати зі словників (або вводити окремо) категорію чи асоціативне поняття, до яких відноситься запит, а також, за необхідністю, поряд вказувати уточнююче поняття або описову інформацію. Чим більше буде лаконічних, необхідних та правильних уточнень метаданих, тим краще пошукова система "зрозуміє" сферу пошуку користувача та сам предмет пошуку, що надасть можливість системі влучно використовувати синоніми для пошуку відповіді саме з цієї області знать.

Отриманий запит пошукова система повинна оброблювати за допомогою семантичного аналізатору. Для цього пропонується використання

нейронних сіток, які здатні налаштовуватися та навчатися за заданими критеріями. В даному випадку такими критеріями будуть слугувати ступені зв’язності конкретних слів – нейронів сітки. Слова-нейрони повинні бути зв’язані декількома способами: синонімічно, категорійно, ієрархічно. Можливі також зв’язки між словами-антонімами. Ці зв’язки повинні бути множинами з нечіткою логікою, які б вказували вагу даної зв’язки. Причина, за якою запропонована така структура сітки, в тому, що за допомогою даних зв’язок нейронна сітка налаштовується семантично. Після того, як сітка навчиться (за допомогою вчителя-користувача, або за допомогою експертів), можна буде за запитом очікувати дійсно релевантну відповідь системи. Релевантність буде досягатися через те, що для кожного зі слів запиту в нейронній сітці будуть активізуватися не лише відповідні слова-нейрони, а й ті, з яким дані слова тісно пов’язані будь-яким зі способів. Тоді формулу (1) можна переписати наступним чином:

$$\text{Релевантність} = TFA * IDFA, \tag{2}$$

де $TFA = f\{k_1 * TF(t_i); k_2 * TF(S_i); k_3 * TF(H_i); k_4 * TF(C_i); k_5 * TF(A_i)\}$,

$$IDFA = \log(|D| / f(k_6 |(d_i \supset t_i)|; k_7 |(d_i \supset S_i)|; k_8 |(d_i \supset H_i)|; k_9 |(d_i \supset C_i)|))$$

де S – синоніми слова у запиті,

H – слова, що пов’язані ієрархічно з словом у запиті,

C – слова, що тісно пов’язані зі словом у запиті за категорією

A – антонім слова у запиті.

$k_1, k_2, k_3, k_4, k_5, k_6, k_7, k_8, k_9$ – постійні числові коефіцієнти.

Згідно даної формули змінюється й підхід до моделі роботи системи пошуку. На рисунку 1 представлена узагальнена модель пошукової системи з використанням нейросіток для семантичного пошуку.

Висновки

Підсумовуючи, можна зазначити, що розглянуті сучасні пошукові системи у багатьох випадках не відповідають необхідним точності та повноті отриманої інформації у відповіді на запит. В основі роботи цих систем лежить синтаксичний пошук та зачатки морфологічного пошуку. Хоча користувачу потрібне виконання наступної задачі: при завдані будь-якої фрази, речення, словосполучення чи просто слова, пошукова система повинна надавати перелік відповідей – незначний, але якісний та точний, своєчасний, саме той, що відповідає його бажанню. Тому постає задача створення нової структури пошукової системи, яка б реалізовувала як синтаксичний, так і семантичний аналіз. Для цього, як один із варіантів, запропоновано реалізувати пошукову систему з використанням нейросітки для збереження слів як нейронів. Ваги зв’язків між словами інтерпретуються як коефіцієнти зв’язків між нейронами. Релевантна відповідь буде та, яка на перетині запиту користувача (разом з синонімами та іншими мета-даними запиту) та документу дасть найбільшу кількість спільних слів-нейронів.

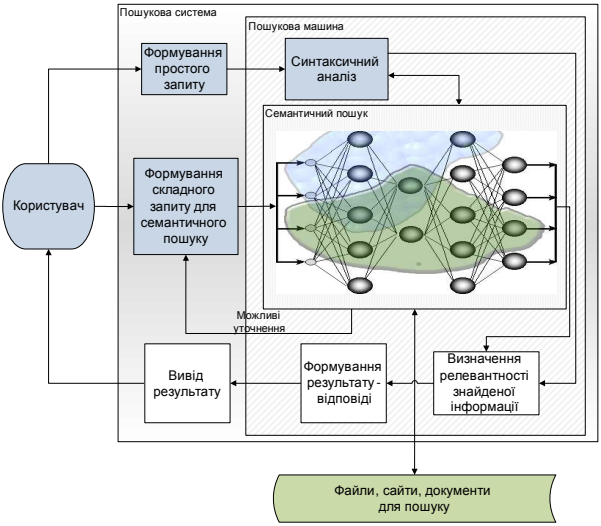

Рис. 1 – Узагальнена модель пошукової системи.

Література

1. Капустин В.А. Основы поиска информации в Інтернет. Методическое пособие. СПб: Институт “Открытое общество” (Фонд Сороса), 1998.
2. Коровенко Ю. Учитывайте особенности русского языка”. URL: http://www.seo-copywrite.ru/52/ (дата обращения 16.03.2010).
3. Текстовое ранжирование в Яндексе, подход TF*IDF. А не изменилась ли формула?”. URL: http://mexboy.ru/?p=126 (дата обращения 01.04.2010).

Отримано 16.02.2010 р.