УДК 681.3

Євтухова Т. І.
ДП Київський державний центр науково-технічної і економічної інформації. Україна, Київ

АНАЛІЗ СУЧАСНОГО СТАНУ ТЕХНОЛОГІЙ ІНФОРМАЦІЙНОГО ПОШУКУ

Анотація

Розглянуто інформаційний пошук в автоматизованій інформаційно-пошуковій системі, який зводиться до трьох основних процесів: перекладання текстів на інформаційну мову; встановлення семантичної відповідності між запитом і документом; знаходження еквівалентних концептів.

Abstract

The information search in the automated information retrieval system is show, which is taken to three basic processes: setting of texts on information language; establishment of semantic accordance between inquiry and document; finding of equivalent kontsepts.

Центральним явищем сучасного етапу процесу побудови інформаційного суспільства є виникнення та розвиток глобальних інформаційних мереж. Об'єднання великого масиву різних документів у цифровій формі за допомогою децентралізованої системи гіпертекстових посилань призвело до створення нових систем інформаційного пошуку, які орієнтовані на роботу в цих глобальних мережах. Висока комерційна ефективність застосування мережевих можливостей стала причиною прискореного широкого впровадження різних пошукових механізмів. Але як результат наукові дослідження часто відстають від потреб практики. Наслідком цього є впровадження розробок, які не цілком адекватні поставленим завданням. Аналізуючи роботу сучасних пошукових систем, можна зробити висновок, що головне їх завдання обмежується тільки видаванням посилань на ті документи, у тексті яких зустрічаються слова з тексту запиту. Але користувача цікавить не це. Йому потрібні не тільки релевантні документи з точки зору співвідношень слів у текстах запиту й у текстах документів, йому цікава сама суть документів, їх смислове значення.

Можна визначити основні причини зниження ефективності пошукових сервісів.

1. Одна з головних задач пошукового сервісу — індексування. Але за індексування відповідає сам сервіс, а не ресурс, якому необхідно індексування.

2. У більшості випадків індексуються окремі сторінки по черзі, а не дані всього ресурсу в цілому.

3. Сучасні пошукові системи не сприймають запити користувача як питання, а вважають, що запити — це сукупність слів, які необхідно знайти в індексі і видати посилання на сайти, де ці слова зустрічаються.

4. Пошукові сервіси видають у відповідь на запити перелік посилань на сайти, але вони не аналізують предметну область цих сайтів, у результаті чого більшість сайтів (інформації на сайтах) просто не відповідають темі запиту.

5. У всіх пошукових сервісах непередбачений засіб, який дозволяв би по запиту користувача видавати необхідну інформацію у вигляді реферату.

6. Сучасні пошукові системи не відслідковують статистику запитів користувачів, у них немає засобів, які б дозволяли аналізувати попередні запити користувачів і на основі цього аналізу видавали б результат.

Загальна схема інформаційного пошуку

Розглянемо основні операції, які необхідні для функціонування ефективного пошукового сервісу.

Першою і найважливішою операцією, пов'язаною з одержанням необхідної інформації, є відбирання документів, що стосуються потрібної теми, — інформаційний пошук. Інформаційний пошук можна визначити як послідовність операцій, що виконуються з метою знаходження документів, які містять певну інформацію (з наступною видачею самих документів або їх копій), або з метою видачі фактичних даних, що являють собою відповідь на поставлені запитання. Пошук текстових документів зводиться до зіставлення двох текстів, з яких один відображає тему, що цікавить споживача, а другий — зміст документа. Перший текст називатимемо запитом, другий — пошуковим рефератом.

Існують два принципово різні підходи до розв'язання завдань інформаційного пошуку — емпіричний і семантичний. В основі першого лежить припущення про те, що інформаційний пошук за своєю суттю є простим процесом, моделювання й автоматизація якого потребує розв'язання лише технічних завдань. Автоматизація зводиться в основному до створення словника термінів з певної галузі знань (словника дескрипторів) та відповідного устаткування для зберігання та пошуку інформації. Другий підхід передбачає, що інформаційний пошук — складний творчий процес, об'єктом якого є зміст. Автоматизація інформаційного пошуку відповідно включає моделювання інтелектуальної

діяльності людини, зокрема розуміння нею змісту текстів.

Розрізняють два види інформаційного пошуку: документальний і фактографічний. Документальна інформаційно-пошукова система у відповідь на запит, у якому сформульовано вимоги до шуканої інформації (наприклад, подаються характеристики певного вузла або пристрою, який цікавить споживача), указує на документи, що містять у собі потрібну інформацію — опис вузла чи пристрою. Фактографічна система видає абоненту (споживачу) безпосередньо шукану інформацію — технічні дані пристрою та ін.

У процесі інформаційного пошуку послідовно розв'язуються дві задачі. Спочатку треба з'ясувати, у яких власне документах міститься шукана інформація, а потім — одержати ці документи або їхні копії.

При інформаційному пошуку запит, як правило, порівнюється не з повним текстом документа, а з його пошуковим образом. Під пошуковим образом документа розуміємо текст, який у стислій формі характеризує основний зміст документа. В автоматизованих інформаційно-пошукових системах пошуковий образ являє собою або перекладені на формалізовану інформаційну мову бібліографічний припис, анотацію чи реферат, або спеціально побудований вираз інформаційної мови, який стисло передає основний зміст документа (реферат). У зв'язку з цим виникає задача автоматизації анотування та реферування.

Процес інформаційного пошуку можна розглянути як ланцюжок послідовних переходів від одного способу або рівня завдання інформації до іншого.

Початковий рівень завдання інформації — речовий. Інформація задається безпосередньо у формі денотата. Денотатом будемо називати клас предметів (тобто множину предметів, що мають деякі спільні властивості або ознаки). Денотат позначимо літерою *d*.

Наступний рівень завдання інформації — логічний. Інформація на цьому рівні задається у формі концептів. Концепт розуміємо як сукупність ознак, що повністю визначають денотат. Ознаки, які утворюють концепт якогось предмета, у своїй сукупності властиві кожному одиничному предметові, що входить у денотат, і не властиві ніякому іншому предметові. При цьому ознакою називається відношення предмета до денотата. Концепт позначимо літерою *c*.

Третій рівень завдання інформації — абстрактно-логічний. Інформація задається у формі імені. Під ім'ям розуміємо адресу денотата, його мітку або номер у певному каталозі, тобто немотивований знак. Ім'ям може бути і слово природної мови, якщо його можна однозначно зіставити з денотатом. Позначимо ім'я через літеру *i*.

Розглянемо таку задачу. Хай на множині концептів *c* задано дві множини функцій *L* та *F*, кожна з яких ставить у відповідність до концепту c_k денотата d_k якийсь концепт $c_k^{'} = c_k$ цього ж денотата, тобто функцій, які здійснюють еквівалентні перетворення концептів (зокрема, може бути $c_k^{'} \supset c_k$ і $c_k^{'} \subset c_k$).

Функція *L* встановлює таку еквівалентність між концептами, яка випливає з даної системи концептів, а функція *F* — еквівалентність, яка обумовлюється, крім того, деякими факторами, що не належать до даної концептуальної системи.

Функція *F* спирається на одиничні зв'язки між денотатами і концептами, тобто зв'язки, закономірність яких або не виявлена взагалі, або не відображена в даній системі. Цілком ясно, що в кожному конкретному випадку можливість застосування функції *L* залежить від того, чи відбиті в системі концептів потрібні зв'язки.

У множині *L* виділяється функція L_1

$$L_1(c_k) = c_k^{'} \subseteq c_n, \qquad (1)$$

і функція L_2

$$L_2(c_k) = c_k^{'} \supseteq c_n, \qquad (2)$$

де $c_k^{'}$ — результат застосування функції *L* до концепту c_k;
c_n — заданий концепт.

У множині *F* виділяється функція F_1

$$F_1(c_k) = c_k^{'} \subseteq c_n, \qquad (3)$$

і функція F_2

$$F_2(c_k) = c_k^{'} \supseteq c_n, \qquad (4)$$

де $c_k^{'}$ — результат застосування функції *L* до концепту c_k;
c_n — заданий концепт.

Надалі функції L_1 та F_1 позначимо f_1, а функції L_2 та F_2 — f_2. Оскільки функції f_1 та f_2 мають смисл тільки тоді, коли задані концепти c_n, включимо c_n в позначення функції таким способом: $f_1(c_k)_{c_n}$ та $f_2(c_k)_{c_n}$.

Тоді мета документального інформаційного пошуку полягає у відборі всіх тих і лише тих імен i_0, концепти яких c_0 задовольняють одну з таких вимог (c_i — концепт запиту):

1) $c_0 \supseteq c_i$;
2) $c_0 \supseteq f_1(c_i)_{c_0}$;
3) $c_0 \supseteq f_1(c_i)_{c_0}$.

Необхідно, щоб концепт кожного відібраного документа (у початковому вигляді або після застосування функцій *L* та *F*) включав концепт запиту (у початковому вигляді або після застосування функцій *L* та *F*).

Таким чином, загальну схему інформаційного пошуку можна відобразити у вигляді такого ланцюжка:

$$c_i' \to l_i' \to s_i' \leftrightarrow s_0' \leftarrow l_0' \leftarrow c_0'$$
$$i_i \to \overset{\uparrow}{c_i} \to l_i \to s_i \leftrightarrow s_0 \leftarrow l_0 \leftarrow \overset{\uparrow}{c_0} \leftarrow i_0, \qquad (5)$$

де i_i, i_0 — ім'я денотата, вказаного в запиті (документі), що надійшов в інформаційно-пошукову систему;
c_i, c_0 — концепт денотата, зазначеного в запиті (документі);
c_i', c_0' — еквівалентний концепт цього ж денотата;
l_i, l_0 — описання концепту природною мовою;
s_i, s_0 — означення концепту інформаційною мовою.

Стрілки показують перехід від одного рівня завдання інформації до іншого, тобто процес її переробки:

$i \to c$ — процес формування концепту для заданого імені (тобто встановлення переліку ознак, які характеризують денотат);
$c \to l$ — процес описування концепту природною мовою;
$l \to s$ — процес перекладу описання концепту з природної мови на інформаційну;
$c \to c'$ — процес знаходження еквівалентного концепту;
$s_i \leftrightarrow s_0$ — процес встановлення відповідності між двома концептами, заданими в термінах інформаційної мови.

У наведеній вище схемі інформаційного пошуку (5) першим рівнем завдання є ім'я. Однак процеси формування концепту та його описування природною мовою здійснюються поза інформаційно-пошуковою системою (якщо не вважати спеціаліста, який формулює запит, одним з її елементів).

Тоді інформаційний пошук в автоматизованій інформаційно-пошуковій системі зводиться до трьох основних процесів: перекладання текстів на інформаційну мову ($l \to s$); встановлення семантичної відповідності між запитом і документом ($s_i \leftrightarrow s_0$); знаходження еквівалентних концептів ($c \to c'$).

Висновки

На сучасному етапі розвитку людства масиви інформації стали вже настільки великі, що пошук потрібної інформації та знань стає вже досить великою проблемою. Для цих цілей розробляються та використовуються різні пошукові механізми, але, як показує досвід, ці пошукові механізмі не завжди відповідають своїй меті. Тому потрібно розробляти нові механізми пошуку.

В роботі отримані такі основні теоретичні та практичні результати:

1. Проаналізовано сучасний стан технологій інформаційного пошуку. Після аналізу стає зрозуміло, що сучасні пошукові механізми досить примітивні, в них відсутні функції інтелектуальності, що погано впливає на релевантність знайдених документів.

2. Розглянуто загальну схему та основні моделі інформаційного пошуку. Інформаційний пошук в автоматизованій інформаційно-пошуковій системі зводиться до трьох основних процесів: перекладання текстів на інформаційну мову; встановлення семантичної відповідності між запитом і документом; знаходження еквівалентних концептів. Переваги та недоліки моделей інформаційного пошуку, за допомогою яких формально описуються ці процеси, не дуже істотні, тому при реалізації пошукового механізму можна використовувати будь-яку модель, а також їх комбінації.

Література

1. *Гринберг План, Гарбер Ли.* Разработка новых технологий информационного поиска // Открытые системы. — 1999. — № 10. — С. 15–30.

2. *Поисковые* системы: http:// meta.math.spbu.ru/ nadejda/irtutorial/ nadejdair.html

3. *Рыбаков Ф. И.* Автоматическое индексирование на естественном языке. — М: Энергия, 1980. — 220 с.

4. *Солтон Дж.* Динамические библиотечные информационные системы. — М: Мир, 1979. — 272 с.

5. *Сэлтон Г.* Автоматическая обработка. Хранение и поиск информации / Пер. с англ. — М: Советское радио, 1972. — 348 с.