УДК 004.738.52

І.Ю. Гришанова[1], С.С. Щербак[2]

[1]*Інститут програмних систем НАН України*
[2]*Харківський національний університет радіоелектроніки, Харків*

РОЗВИТОК ТЕХНОЛОГІЙ ІНФОРМАЦІЙНОГО ПОШУКУ ТА АНАЛІЗ ЇХ ЗАСТОСУВАННЯ В SEMANTIC WEB

Аналізуються технологічні аспекти впровадження пошукових засобів в розподілене середовище Semantic Web. Надаються адаптовані до впровадження в Semantic Web концепції інформаційного пошуку, специфікації задач, моделі та класифікація пошукових систем за різними ознаками. Наводиться огляд існуючих на даний час пошукових систем та надається перелік ознак семантичних пошукових систем.

Ключові слова: *Семантик Веб, таксономія, інформаційний пошук, онтології, пошукові механізми.*

Вступ. Роль персоніфікації в процесі пошуку

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers) [1].

Класичне поняття ***Інформаційного пошуку*** (IR – information retrieval, ІП) базується на задоволенні інформаційної потреби користувачів (information need), тобто потреби в пошуку релевантної інформації. Загальноприйняте визначення інформаційного пошуку базується на необхідності користувачеві отримати певну інформацію.

Під ***інформаційною потребою*** зазвичай розуміється виражена в інформаційному запиті необхідність в інформації, яка повинна бути задоволена. Наприклад, планування поїздки формує інформаційну потребу вивчити розклад руху потягів та іншого транспорту. Такий процес може бути виконаний різними способами: за допомогою телефону, в агентстві з продажу квитків, безпосередньо в касах вокзалу, за допомогою веб-сторінки перевізника або через пошук в Інтернеті. Однак, незалежно від форм задоволення інформаційної потреби, вона не змінюється.

Необхідно зазначити, що коли потрібний маршрут вже обрано, а квитки придбані, для користувача ця інформація втрачає свою цінність, при цьому вона залишається цінною для інших потенційних споживачів. Така властивість повної втрати цінності інформації (її споживацької вартості) для певного споживача в деякий момент часу, є важливою особливістю інформаційної потреби, що суттєво відрізняє її від інших видів потреб людини. Одна й та сама інформація знову може стати предметом споживання, якщо вона буде надана іншому споживачеві, або якщо перед тим самим споживачем знову стане таж сама задача, або якщо запас знань споживача зросте, що дозволить йому побачити в цій інформації нові аспекти.

Таким чином, інформаційні потреби носять суто індивідуальний (персональний) характер. Вони залежать не тільки від особливостей задач, що вирішуються, але й від психологічних, освітніх та інших особистих рис особи, що приймає рішення.

Розрізняють два основних типи інформаційних потреб:

– поточні, які зумовлені притаманною людині допитливістю, і які виражаються в його прагненні бути в курсі всього, що відбувається в світі;

– конкретні (спеціальні), які виражаються в прагненні отримати інформацію, необхідну для вирішення конкретної задачі – дослідницької, професійної, управлінської тощо [2].

Основна мета інформаційного пошуку – допомогти користувачеві знайти інформацію, яка йому необхідна. Процес інформаційного пошуку в загальному вигляді включає в себе послідовність операцій, направлених на збір, обробку та надання необхідної інформації зацікавленим особам. Процес інформаційного пошуку складається з наступних етапів:

– визначення (уточнення) інформаційної потреби та формулювання інформаційного запиту;

– визначення сукупності можливих інформаційних джерел;

– вилучення інформації з виявлених інформаційних джерел;

– ознайомлення з отриманою інформацією і оцінювання результатів пошуку.

Базовими поняттями оцінювання ефективності пошуку є релевантність.

Вирішальною умовою ефективного задоволення інформаційної потреби є чітке усвідомлення і чітке вираження того, яка інформація насправді потрібна споживачеві для вирішення поставленої перед ним задачі. Без цього складно розраховувати на отримання релевантного результату.

З моменту виникнення у людини інформаційної потреби, вона починає оцінювати всю інформацію, що надходить до неї, під кутом зору цієї потреби, розділяючи цю інформацію на релевантну та нерелевантну. Іншими словами, інформаційна потреба виникає у людини при постановці перед нею певної задачі, під час обмірковування якої, в людини складається образ або модель даної задачі. Цей

образ і служить еталоном, з яким порівнюється вся подальша інформація, що надходить. Якщо інформація має відношення до еталону, вона вважається доречною. Все, що не має відношення до еталону – вважається нерелевантною інформацією.

Під впливом міркувань над сутністю поставленої задачі та вмістом релевантної інформації, що накопичується, уява людини про цю задачу може уточнюватися та змінюватися. Психологи називають такий процес зростанням стану поінформованості про задачу.

Коли людиною накопичено необхідну кількість інформації і виконано деякий міркувальний процес, вона знаходить розв'язок задачі. Після цього вся інформація, пов'язана з розв'язком задачі, переміщується в зону архівного зберігання. Таким чином, інформаційна потреба може бути охарактеризована як усвідомлена потреба в інформації, необхідної для вирішення за розробленим планом поставленої задачі.

Можна припустити, що процес вирішення будь-якої наукової задачі починається з прийняття різних передумов і припущень, які в подальшому будуть коригуватися та змінюватись. Під образом чи моделлю задачі слід розуміти гіпотезу, яка є важливим засобом організації наукового пошуку.

Вчення про психологічні установки дозволяє пояснити поняття пертинентності (або релевантності), яке є одним з ключових понять теорії інформаційного пошуку. Під пертинентністю розуміється відповідність знайдених документів або відомостей справжній інформаційній потребі вченого або спеціаліста, яку він нерідко сам може ясно не усвідомлювати.

З запропонованої інтерпретації змісту інформаційної потреби та механізму її задоволення випливає, що віднесення інформації, що надходить до людини, до категорії релевантної чи нерелевантної, повністю визначається тим, який образ поставленої задачі склався у даної людини. Сам цей образ залежить, принаймні, від трьох наступних факторів:

– інформації, яка вже накопичена людиною в її пам'яті;
– обраного методу розв'язання задачі;
– темпів і проміжних результатів розв'язку.

Ще раз необхідно зазначити, що образ задачі під впливом інформації, що надходить, та проміжних результатів розв'язання цієї задачі, уточнюється або навіть змінюється. У зв'язку з цим змінюються і ознаки, за якими розпізнається і відбирається релевантна інформація. Тому для адекватного інформаційного обслуговування фахівців необхідно, щоб процес пошуку був не тільки індивідуальним, але й включав в себе постійний зворотній зв'язок для своєчасного врахування змін в їх інформаційній потребі.

Базові поняття інформаційного пошуку

Основним засобом передачі інформації в часі та просторі є документ, який можна визначити як засіб збереження певної інформації про факти, події, явища, об'єктивну дійсність та розумову діяльність людини на спеціальному матеріалі [3]. Документи мають різну форму представлення. В автоматизованих інформаційно-пошукових системах це текстова інформація на природній мові. В повсякденному житті – це може бути друкована стаття, книга тощо. В Інтернет це може бути малюнок, відео-ролик або веб-сторінка.

Поняття інформаційного пошуку (ІП) в інформатиці вперше було запроваджено в 1947 році американським математиком Келвіном Муерсом. Інформаційним пошуком називається деяка послідовність операцій, яка виконується з метою знаходження документів, що містять певну інформацію (з подальшою видачею цих документів або їх копій), або з метою видачі фактичних даних, які дають відповіді на задані питання [4].

Як було зазначено вище, до інформаційного пошуку спонукає інформаційна потреба, виражена у формі інформаційного запиту. Об'єктами інформаційного пошуку можуть бути документи, відомості про їх наявність та/або місцезнаходження, фактографічна інформація.

Інформаційний запит представляє собою сформульовану на природній мові інформаційну потребу. Результат «перекладу» інформаційного запиту на інформаційно-пошукову мову (ІПМ) називають ***пошуковим образом запиту*** (ПОЗ). Синтаксис і семантика ІПМ визначається структурою та наповненням документів, а також загальними задачами системи.

Інформаційний пошук розрізняють:

– в залежності від мети – адресний пошук (формально-механічний) та семантичний (тематичний);
– в залежності від об'єкту пошуку – документний та фактографічний;
– в залежності від ступеню використання технічних засобів – ручний або автоматизований.

Всі види інформаційного пошуку перетинаються, тому що цілі та об'єкти часто взаємопов'язані. Наприклад, документний і фактографічний види пошуку можуть бути як адресними, так і семантичними.

Найбільш ефективним методом пошуку документів, що містять інформацію, є вивчення (прочитання) кожного окремого документу. Зрозуміло, що такий спосіб практично неможливий, оскільки кількість документів, як правило, буває занадто великою, щоб можна було прочитати всі документи при кожному інформаційному запиті[5]. В звязку з цим необхідна розробка ефективних технологій та засобів пошуку інформації.

Види пошуку в веб-середовищі

Поява та розвиток Інтернет сприяли розширенню поняття пошуку та появі більш специфічного поняття – веб-пошуку. Оскільки в контексті Веб фактори взаємодії людини з комп'ютером та когнітивні аспекти грають найважливішу роль[6], корисно деталізувати модель інформаційного пошуку, як це показано на рис. 1.

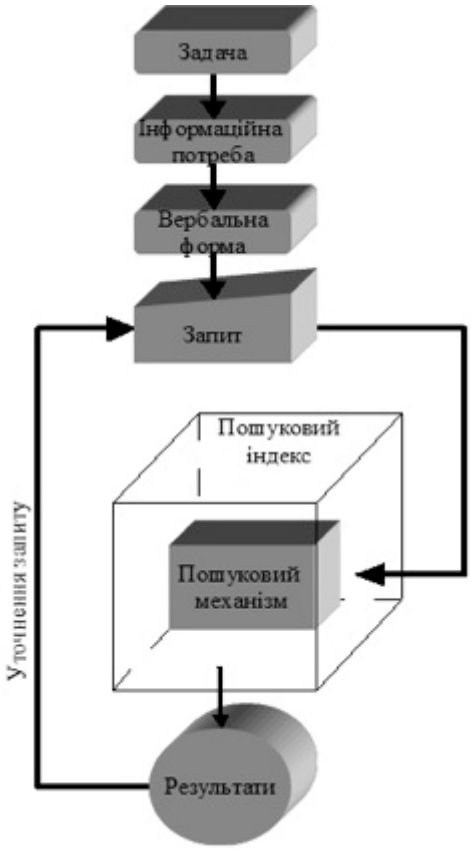

Рис. 1. Модель інформаційного пошуку, поширена на Інтернет-мережу (веб).

Як було зазначено раніше, інформаційна потреба асоціюється з певною задачею. Ця потреба вербалізується (найбільш часто це виконується ментально та не дуже чітко) та транслюється в запит, що надається пошуковому механізму. Цей процес висвітлення та створення запиту з інформаційної потреби здобув в контексті Веб велику увагу: в статті Хольстера та Струбе [4] вказується, що досвідчені користувачі та користувачі-початківці конструюють запити по-різному. Наварро-П'єтро та ін. [5] вивели когнітивну модель для веб-пошуку, Мурамату та Прат [6] дослідили ментальну модель користувачів пошукових механізмів, тощо (див. також [7]). Однак приведені дослідження базуються на припущенні, що користувачі пошукових веб-систем мотивовані інформаційною потребою.

З іншого боку, в контексті Веб, «потреба, викликана запитом» часто не є інформативною.

В роботі [8] приведено класифікацію запитів у відповідності до їх спрямованості:

– навігаційні запити. Такі запити мають на меті негайний намір побачити певну веб-сторінку;

– інформаційні запити. Вони виражають намір одержати деяку інформацію, яка вважається існуючою на одній або більше веб-сторінках;

– транзакційні запити. Ці запити виражають намір виконати якусь веб-опосередковану діяльність – покупку в інтернет-магазині, завантаження файлів тощо.

Для виконання приведених типів запитів використовуються сучасні пошукові системи, які добре виконують інформаційні та навігаційні запити, але транзакційні запити виконуються лише опосередковано, тобто результат їх виконання нерелевантний. Шлях підвищення ефективності пошуку лежить в удосконаленні семантичного аналізу (тобто розуміння змісту запиту) та змішування різних зовнішніх баз даних.

Визначення пошуку в Веб-середовищі. В зв'язку з появою Веб, поняття пошуку в середовищі Інтернет набуло іншого змісту. Поняття пошукової системи (ПС) стало більш широким та глибшим. Наведемо декілька нових визначень поняття пошукової системи (Search Engine), що використовуються зараз в західній науковій літературі.

Пошукова система – це комп'ютерна програма, яка отримує (retrieves) файли або документи, або дані з бази даних або з комп'ютерної мережі (зокрема, з Інтернету) [9]. Пошукова система – це комп'ютерна програма, яка знаходить (finds) інформацію в Інтернеті шляхом пошуку слів, які були введені (як запит – прим. автору) [10]. Пошукова система – це комп'ютерне програмне забезпечення для пошуку даних (з текстів або баз даних) для отримання конкретної інформації, а також веб-сторінка у веб-мережі, яка використовує таке програмне забезпечення для пошуку ключових слів на інших сайтах [11].

В контексті Веб з огляду на тезу, що «потреба спонукає запит», в клас поняття пошукових систем почали включати системи «питання-відповідь» (answer engine), які дуже часто є фактографічними ПС. Але деякі системи для отримання результату пошуку вже починають використовувати процедури логічного виведення.

Таким чином, в контексті Веб пошукова система (а точніше – інформаційно-пошукова система, ІПС) використовує спеціалізоване програмне забезпечення, яке має на вході від користувача пошуковий/і термін/и і на виході надає список веб-сторінок, які вважаються найбільш релевантними. Більшість пошукових систем мають величезні бази даних мільярдів веб-сторінок. Розрізняють два типи веб-пошукових систем.

Пошукові системи, що базуються на кроулінгу (Crawler-based). Такі системи створюють свої списки веб-сторінок автоматично. Вони «сканують» (crawl) Інтернет за допомогою робота-«павука» (spider, програма, яка відвідує веб-сторінки, читає їх і переходить далі за посиланнями, знайденими на веб-сторінці), та повертають користувачу результати пошуку, ранжовані у порядку важливості. Павук повторно відвідує веб-сторінки кожні кілька місяців для оновлення своєї індексної бази відповідно до внесених на веб-сторінки змін. Головна перевага пошукових систем, базованих на кроулінгу, полягає у тому, що будь-які зміни, внесені до веб-сторінки, будуть впливати на його базу і відповідно на результати пошуку. Таким чином, актуальність змісту веб-сторінок збігається з ключовими словами, що використовуються для пошуку.

Каталоги, створені людиною (human based directory), залежать від людей, що їх створили та поповнюють. Вони виконують пошук за ключовими словами в коротких описах веб-сторінок, представлених веб-майстерами та спеціалістами, що рецензують та перевіряють каталог. Разом з цим, веб-сторінки переглядаються людиною і розміщуються у відповідну ієрархію категорій. Таким чином, зміни, внесені до веб-сторінки, на відміну від скануючих пошукових систем, не будуть мати ніякого впливу на збережений в каталозі опис. Отже, хоча на веб-сторінці й міститься відповідна інформація, що відповідає запиту, але вона не буде відображена в списку результатів пошуку доки веб-майстер не змінить опис веб-сторінки. Саме з цієї причини один з найперших та найбільших каталог, сформований людиною Yahoo! перетворено у більш популярну пошукову систему на базі сканеру. Таким чином утворюються комбіновані пошукові системи. Оскільки каталоги містять інформацію, перевірену людиною, ця інформація використовується для фільтрування та ранжування результатів пошуку.

Типи пошукових механізмів:

- пошукові системи;
- веб-каталоги;
- віртуальні бібліотеки;
- мета-пошукові механізми.

Пошукові системи (Search Engines) є найбільш широким класом ІПС та найбільш популярним і загальновживаним. Вони характеризуються наступними властивостями: мають базу даних веб-сторінок; пошук здійснюють за ключовими словами; мають скануючого робота. Яскравим прикладом такої системи є пошукова система Google.

Веб-каталоги (Web Directories). Як було вказано вище, вони мають колекцію веб-ресурсів; організовані за тематичними категоріями в ієрархію; організація в категорії та інш. робиться вручну. Приклад такого каталогу – загальновідомий каталог Yahoo.

Віртуальні бібліотеки (Virtual Libraries). Такі бібліотеки характеризуються наступними ознаками: мають колекцію веб-джерел; оцінюються фахівцями з предметної області; слабо автоматизовані, живляться людськими ресурсами. Приклад типової бібліотеки – бібліотечний індекс інтернету – Librarians Index to the Internet www.lii.org.

Мета-пошукові механізми (Meta-Search Tools). З назви видно, що такі механізми використовують ресурси інших пошукових систем, а результати фільтрують та ранжують згідно своїх заданих правил. Такі системи мають наступні властивості: не мають власної бази даних; вони здійснюють запити до інших пошукових механізмів, розташованих в Веб; мають дуже поганий дизайн і можуть тільки змінювати порядок ранжування результатів. Класичний приклад такої системи є MetaCrawler.com. Такі системи користуються попитом, оскільки вони повертають більш короткий список посилань, що психологічно більш прийнятно для людини.

Еволюція пошукових систем Інтернет

Перше покоління пошукових систем використовувало в основному інформацію, яка знаходилась безпосередньо на веб-сторінках (текст і форматування), ці пошукові системи дуже близькі до класичних ПС [**Ошибка! Закладка не определена.**]. Такі системи виконують в основному тільки інформаційні запити. Типовими прикладами таких систем в 1995 – 1997 роках були загальновідомі AltaVista, Excite, Webcrawler і т.д. Ранжування сайтів відбувалося тільки за рахунок контенту сторінок.

Важливі фактори, які враховувалися при ранжуванні, включали щільність ключових слів на веб-сторінці, назву і місцезнаходження цих ключових слів у документі. Також ПС першого покоління для обчислення релевантності враховували мета-теги, використання ключових слів в імені домену, а також в url-адресі (докладніше – див. [12]).

Основні спам-фільтри цих систем робили перевірку на наявність ключових слів в тексті, представлених на сторінці тим самим кольором, що і фон документу, тобто які невидимі людському зору. На той час з'явилися перші портали, в наслідок чого результати пошуку перетворилися в величезні рекламні щити та перевантажені інформацією жовті сторінки.

Друге покоління пошукових систем (початок появи 1998 – 1999 р.) характеризується використанням інформації, яка існує поза веб-сторінкою, – веб-специфічних даних, таких, як аналіз посилань (link analysis) та відстеження даних, що передаються з http-запитом (click-through data). Таким чином вони почали враховувати структуру веб-мережі.

Друге покоління більш щільно пов'язано з семантикою запитів, яка береться з аналізу даних, що подані в Веб поза сторінки. Деякі з основних компонентів, які вони використовують, є відстеження кліків (tracking clicks), репутація сторінки (page reputation), індекс популярності (link popularity), темпоральні спостереження (temporal tracking, кількість часу, що проводять відвідувачі на сторінці), та якість посилань (link quality). Пізніше, ПС другого покоління почали використовувати вектори термів (term vectors) [13], аналіз статистики відвідування (stats analysis), кеш-дані (cache data) і контекст. В якості аналізу контексту розглядається пошук на сторінці пар ключових слів, які складаються з двох слів. Це дозволяє краще виконати віднесення сторінки до певної категорії.

Першою системою, яка почала використовувати аналіз посилань між сторінками в якості одного з основних факторів ранжування, стала система Google (PageRank). ПС DirectHit стала першою, хто побудував ранжування на аналізі даних, що передаються під час http-запиту. На сьогоднішній день всі основні системи використовують приведені типи даних. Використання Google PageRank, методу відстеження кліків DirectHit та тривалості візиту підвищило ефективність пошуку.

Пошукові системи другого покоління підтримують як інформаційні, так і навігаційні запити. Аналіз посилань мають вирішальне значення для навігаційних запитів.

На даний час зароджується третє покоління пошукових систем. Ці пошукові системи є спробою поєднати дані з різних джерел для досягнення головної мети – видачі результату, що відповідає потребі користувача. Наприклад, на запит «Київ», ПС повинна надавати пряме посилання на сторінку бронювання готелів в Києві, сервер мап з мапою міста, на сервер погоди з інформацією про погоду і т.д. Таким чином, третє покоління – це покоління пошукових систем, які виходять за рамки обмежень фіксованої бази даних за допомогою семантичного аналізу, визначення контексту пошуку, вибору динамічної бази даних і т.д. Задача полягає в тому, щоб забезпечити інформаційні, навігаційні та транзакційні запити.

Третє покоління пошукових технологій покликані об'єднати масштабованість існуючих Інтернет-пошукових систем з новими та вдосконаленими моделями пошуку релевантності; вони починають враховувати переваги користувача, співробітництво, колективний інтелект, багатий досвід користувачів та багато інших спеціалізованих можливостей, які роблять інформацію більш вагомою, а пошук – більш продуктивним.

Пошукові системи третього покоління додають до бази даних векторів термів похідні слова (word stemming) і тезаурус, що надає допомогу у здійсненні пошуку за контекстом [14]. Автоматичне визначення ключових пар також допомагає автоматичній категоризації сторінки, визначенню де користувач хоче провести покупку, а де – здійснити пошук, що повинно видати абсолютно різні результати пошуку на основі контексту або намірів користувача.

Технології третього покоління збагачені картами Веб, які є корисними для фільтрації – видалення дублікатів сайтів, а також багатьох самостійних сторінок, які привертають трафік на всього лише декілька ключових слів. Це означає, що сторінки типу дорвеїв (doorways), гейтвеїв (gateways), вхідних (entry, splash) – спеціально створені спам-сторінки для цільової розкрутки сайту на визначені позиції ключових слів, – незабаром будуть відфільтровані.

Вони також будуть вилучати як можна більше даних про індивідуальні пошукові звички користувача. Всі основні пошукові системи планують створення персональних профілів та агентів, які будуть накопичувати знання про користувача протягом певного періоду часу та використовувати їх, виходячи з минулих пошукових звичок.

Поява Семантичного Вебу (докладніше див. [14]) надало нові можливості і ще більше диференціювало поняття інформаційного пошуку. Семантичний Веб надав можливість використовувати існуючу семантичну інформацію – представлену за допомогою семантичної розмітки, використовуючи семантичні зв'язки, що виконують різні операції виведення на семантичних даних, а також порівняння семантичної інформації. Змінюється й алгоритм ранжування результуючих документів – вводиться поняття семантичного ранжування документів. Змінюється алгоритм пошуку, він стає дедалі розподіленим, змінюються методи представлення пошукового запиту. Поява різних типів поданої в Веб інформації (різної модальності – мультимедійної інформації, відео, аудіо тощо) потребує використання інших підходів до пошуку. Існуюче розділення пошуку за типом інформації – пошук відео, пошук картинок, тощо (Google, Яндекс) – дуже стиснене та неінформативне. Існує синергетична потреба – виконання пошуку в різних типах інформації та подальше змішування результатів.

Таким чином, пошукові системи третього покоління виходять за рамки класичного (традиційного) поняття пошуку. Це пов'язано з появою нових типів інформації та нових вимог, що ставлять користувачі перед пошуковими системами.

В західній літературі з'явився термін Search 2.0, який асоціюється з третім поколінням, але має більш чіткі риси, та орієнтований на бізнес-аудиторію. [15]. У Веб вже існує ряд проектів, які вважаються проектами search 2.0 – Swicki, Rollyo, Clusty, Wink, Lexxe тощо.

Приклади технологічних рішень пошукових систем третього покоління

З розвитком нових технологій та стандартів, паралельно з науковими дослідженнями, та опираючись на них, компанії бізнес-сектору прагматично розвивають нове покоління пошукових систем – семантичних (розумних) ПС, «smarter» search engines. Наведемо приклади таких технологічних рішень пошукових систем, які інтелектуалізують процес пошуку за рахунок:

– структурування та представлення даних, отриманих з Інтернет;

– реалізації семантичної фільтрації за якістю;

– організації пошуку серед структурованих даних в Інтернет;

– пошуку в режимі реального часу в Інтернет;

– пошуку в «глибинному» Веб («deep web») [16].

Структурування та подання впроваджують Wolfram Alpha, Google Squared, Sensebot.

Wolfram Alpha – це система обчислення знань, яка почала працювати 5 березня 2009 року. Її засновником є британський фізик Стівен Вольфрам, голова компанії «Wolfram Research» та розробник широко відомої в наукових колах програми Mathematica.

«We aim to collect and curate all objective data; implement every known model, method, and algorithm; and make it possible to compute whatever can be computed about anything. Our goal is to build on the achievements of science and other systematizations of knowledge to provide a single source that can be relied on by everyone for definitive answers to factual queries.» – Stephen Wolfram [17].

Деякі називають «Wolfram Alpha» пошуковою системою, але на відміну від традиційних пошукових систем, які обмежуються тим, що за запитом користувача видають список посилань на веб-сторінки, що мають відповідати запиту, – сервіс «Wolfram Alpha» самостійно аналізує запити користувача і представляє йому зведену релевантну інформацію.

З огляду на прийняту класифікацію ця система є системою «запит-відповідь». Автор позиціонує систему не як пошукову (search engine), а як Computational Knowledge Engine («Систему Обчислювального Знання»), він каже: «Наша мета – зробити знання доступними всім, коли завгодно і де завгодно».

Ця система об'єднує обчислювальні потужності Mathematica з інструментами, які експліцитно оперують з усіма типами даних. Дані повинні негайно оброблятися, щоб надати можливість «брати питання людей, виражені природною мовою, і представляти їх у точній формі, яку можна буде обробляти, а також керувати всіми нотатками, зробленими людьми з усіх можливих предметних галузей» [17].

IT-аналітики вже охрестили Wolfram Alpha «інтелектуальною пошуковою системою», «пошуковою системою нового покоління», «інтернет-генератором розумних відповідей».

Спочатку Wolfram Alpha працював у закритому (тестовому) режимі, а з 18 травня 2009 р. веб-сервіс вже відкритий для всіх бажаючих. За час закритого тестування було оброблено близько 23 млн. запитів, а за перший тиждень після відкриття – близько 100 млн. На сьогоднішній день Wolfram Alpha є безкоштовним веб-сервісом. Надалі планується розміщувати на сторінках проекту рекламу, а також пропонувати користувачам професійну версію з додатковими функціями за невелику плату.

Предметні області, які обробляються в системі – математика, фізика, хімія, астрономія, статистика та різні дані статистичного аналізу, дати та час, географія, погода, здоров'я та медицина, культура та медіа, музика та освіта, люди та історія, фінанси, лінгвістика і досягнення високих технологій, спорт тощо.

Система Wolfram Alpha дозволяє [18]:

– переводити одиниці виміру з однієї системи в іншу;

– якщо задати в якості запиту хімічну формулу, система видасть основну інформацію про цю речовину/хімічний елемент;

– якщо ввести в рядок пошуку 1 apple + 1 orange, – система видасть кількість калорій, протеїнів, вітамінів, відсутність/наявність холестерину і т.д.;

– якщо ввести назву міста, то система видає наступну інформацію: де воно знаходиться, кількість жителів, схематичне розташування на карті, поточний час, поточну температуру, вологість, швидкість вітру, стан хмарності, висоту над рівнем моря, найближчі міста (з відстанню до них і з кількістю мешканців у цих містах). Натиснувши на посилання «Show coordinates», можна дізнатися координати міста. Натиснувши на посилання «Satellite image», система завантажить знімки свого міста (буде завантажена веб-сторінка «Карти Google»);

– система виконує різні обчислення: якщо ввести в рядок пошуку, наприклад, $ 999 + 15%, Wolfram Alpha зробить необхідні обчислення;

– система надає інформацію про будь-яку веб-сторінку. Якщо ввести в рядок пошуку URL веб-сторінки, система видасть детальну інформацію: хто є хостинг-провайдером, де він розташований, кількість переглядів і кількість візитерів за добу, site rank, найменування та розмір титульної сторінки, кількість вихідних посилань, кількість «зображень»;

– система може проводити не тільки найпростіші обчислення, але й розв'язувати різні рівняння: якщо ввести, наприклад, x ^ 3 sin (x), система видасть розв'язок у вигляді графіка та в аналітичному вигляді;

– обробка музики – якщо ввести в рядок пошуку, наприклад, C Eb GC (символьні позначення нот), то система надасть вичерпну інформацію про ці музичні ноти;

– обробка імен – якщо ввести два різних імені, наприклад, Vera, Natasha, - в результаті система видає статистичні дані, що свідчать про те, як часто використовуються ці імена;

– обробка фінансової інформації: система може надавати інформацію про економічний стан (наприклад, про наявність акціонерного капіталу, вартості однієї акції і т.д.) двох компаній, назви яких вводяться у пошуковий рядок з пробілом між назвами;

– обробка часової інформації: якщо ввести дату у форматі, наприклад, august 28, 1959, то система видасть, який це був день тижня, можна буде підрахувати, скільки часу (років, місяців, тижнів, днів) пройшло з цієї дати, хто з відомих людей народився в цей день, які свята припадають на цей день.

Для того, щоб дізнатися джерела інформації, які використовував Wolfram Alpha, внизу, під знайденою інформацією знаходиться кнопка «Source information».

Всю інформацію, яку згенерував («навольфрамив» – сленг) Wolfram Alpha, можна зберегти у вигляді PDF-файлу, натиснувши посилання (внизу) «Download as: PDF».

На жаль, система обробляє тільки англомовні запити.

Google Squared – цей новий експериментальний пошуковий механізм (experimental search tool) було заявлено 3 червня 2009 р. На відміну від класичних – «традиційних» – пошукових систем, Google Squared не видає на запит користувача сторінку зі списком посилань на веб-ресурси, що відповідають запиту. В якості результатів пошуку на екран користувача виводиться зведена таблиця з інформацією по запиту. Тобто Google Squared, як і сервіс Wolfram Alpha, самостійно аналізує (намагається аналізувати) запити користувача і представляє йому зведену релевантну інформацію.

В офіційному блозі пошукового гіганта сказано так: «...Іноді знайти інформацію легко. А іноді для збору необхідної інформації потрібно відвідати 10-20 веб-сторінок, а потім ще потрібно структурувати знайдене ... Squared Google не шукає веб-сторінки за вашим запитом. Замість цього, він автоматично вибирає й організовує факти зі всього Інтернет» [19].

Google Squared дозволяє керувати вмістом результуючої зведеної таблиці: можна додавати до таблиці нові рядки і стовпчики, а зайві – видаляти. Крім того, можна додавати до зведеної таблиці дані з нового пошукового запиту.

Оскільки інформація збирається з різних джерел, для однієї комірки таблиці Google Squared може знайти відразу кілька значень. Відображається при цьому тільки одне значення, але користувач може при бажанні вибрати інше.

Як і Wolfram Alpha, сервіс Google Squared не підтримує українську та російську мови.

Порівняльне тестування Google Squared та Wolfram Alpha, наведене в [20], показує, що аналітичні характеристики і можливості системи Google Squared на даний час явно поступаються Wolfram Alpha.

Оскільки ця система є комерційною, то знайти на даний час більш детальний опис алгоритмів та принципів роботи, нажаль, майже неможливо.

SenseBot заявлена як семантична пошукова система, яка на пошуковий запит генерує текстові анотації (резюме), складені з веб-сторінок, що відносяться до теми пошукового запиту. Ця система для вилучення змісту з веб-сторінок і представлення його користувачеві узгодженим чином використовує інтелектуальну обробку текстів (text mining) і мультидокументну сумаризацію (multidocument summarization). Разом з результатами система видає «семантичну хмару» концептів («Semantic Cloud» of concepts), що дозволяє направити увагу користувача системи та керувати результатами.

Оскільки SenseBot є семантичною пошуковою системою, це означає, що вона намагається зрозуміти семантику отриманих в результаті сторінок. Вона використовує, як було зазначено вище, інтелектуальну обробку текстів для розбору web-сторінок і визначення їх основних семантичних концептів. Після цього вона виконує мультидокументну сумаризацію змісту, в результаті чого генерується зв'язане резюме.

На верхньому рівні, система отримує джерела, які видаються як результат основною пошуковою системою. Після цього система виконує інтелектуальну обробку тексту, отриманого з кожного джерела, вилучаючи ключові концепти. Між джерелами оцінюються подібності, і ті, що семантично знаходяться далеко від запиту, або не зв'язані з загальною масою знайдених джерел, відкидаються. Концептам присвоюється вага, а також задається преференційні значення для концептів, які представлені у запиті. Потім виконується відповідно до запатентованого алгоритму мультидокументна сумаризація – збір з текстів резюме (зі знайдених документів) остаточного документу. Таким чином, на запит користувача фактичними результатами веб-пошуку є резюме, згенероване зі знайдених документів.

Найкращі результати можуть бути досягнуті на множині текстових документів, які по суті знаходяться близько до заданої теми. Найкраща область застосування цієї системи, як зазначає її розробник, є вертикальні пошукові системи й портали – фінансові, медичні, правові, бібліотеки і т.д. Що стосується загального веб-пошуку, деяка кількість «шуму» неминуча, навіть для тих джерел, що знаходяться на перших сторінках результатів, які вважаються найбільш релевантними [21].

Проектом «Hakia» запропонована семантична фільтрація інформації за якістю. Цей проект розвивається вже декілька років, але до теперішнього часу знаходиться в стадії бета і добре охоплює поки що предметну область з медицини та здоров'я. Семантична технологія «Hakia» забезпечує новий досвід пошуку, який орієнтований на якість, а не популярність. Якісні результати повинні задовольняти трьом критеріям одночасно:

– вони надходять з заслуговуючих довіри вебсторінок, рекомендованих бібліотекарями;

– представляють собою найбільш свіжу наявну інформацію;

– залишаються абсолютно релевантними до запиту.

Організацією пошуку серед структурованих даних у Веб займаються в SWSE та Swoogle.

На даний час вже існує багато даних, які відповідають запропонованим стандартам семантичного вебу (наприклад RDF та OWL). Вже існує багато малих вертикальних словників і онтологій, які все більше використовуються різними спільнотами для досягнення своїх цілей. Користувачі вебу публікують описи своїх профілів, з використанням FOAF (Friend of a Friend), провайдери новин транслюють добірку новин в вигляді RSS (RDF Site Summary), зображення анотуються з використанням різноманітних RDF-словників.

SWSE представляє собою сервіс, який постійно вивчає та індексує семантичний веб (Semantic Web) і забезпечує легкий у використанні інтерфейс, за допомогою якого користувачі можуть знайти дані, що їх цікавлять.

SWSE індексує триплети RDF або OWL, знайдені у Веб, і надає послугу з пошуку серед цих триплетів.

Swoogle також є пошуковою системою, створеною для семантичного веб. Роботи Swoogle сканують веб з метою пошуку спеціального класу вебдокументів, які називаються семантичними вебдокументами, тобто які написані в RDF. Ця пошукова система також виконує пошук серед RDFтриплетів і видає посилання на джерела, які їх містять. Пошук здійснюється за ключовими словами.

Аналогічні функции пропонують пошукові системи WatsOn, Semanticwebsearch, Sindice та Falcons.

Пошук в режимі реального часу в Веб реалізують OneRiot та Scoopler.

OneRiot сканує посилання, якими користувачі діляться на сервісах посилань (міток) Twitter, Digg та інших соціальних сервісах, а потім індексує зміст цих сторінок. Таке індексування пошукова система проводить в режимі реального часу – поява нового посилання на сервісі одразу викликає процес індексування. Кінцевим результатом роботи пошукової системи є пошуковий досвід, який дозволяє користувачам знаходити свіжий, найбільш соціально вагомий контент в реальному часі у Веб. Результати пошуку індексуються в залежності від їх актуальності та популярності.

Scoopler – це пошукова система, яка виконує пошук в режимі реального часу. Робот цієї пошукової системи збирає та організовує для загального користування контент, по мірі його виникнення у Веб. Таким контентом ця система вважає доповіді головних новин, фотографії та відеоматеріали значних подій, а також посилання на найгарячіші нотатки поточного дня. Джерелами контенту, який індексується, є постійні оновлення, що надходять з сервісів Twitter, Flickr, Digg, Delicious тощо.

Однак, найбільш цікавим є пошук в «глибинному» Веб («deep web»). Однією з систем, які реалізують такій пошук є DeepDyve.

DeepDyve – «пошуково-дослідницька» система, яка використовує власні (комерційні) технології пошуку та індексування, що дозволяють відбирати багатий, релевантний контент з тисячі журналів, мільйонів документів і мільярдів незадіяних веб-сторінок глибинного вебу. Дослідники, студенти, технічні спеціалісти, бізнес-користувачі, а також споживачі іншої інформації, можуть отримати доступ до багатої незадіяної інформації, що зберігається в «глибинному» Веб – інформації, яка займає переважну більшість Інтернет та не індексується традиційними пошуковими системами. Пошуково-дослідницька система DeepDyve відчиняє шлях до цього поглибленого професійного контенту і повертає результати, які не навантажені інформацією з оглядових (реферативних) сайтів і нерелевантною інформацією.

Система використовує запатентований алгоритм KeyPhrase™, який застосовує метод індексації, отриманий при дослідженнях в області геномікі. Алгоритм шукає збіг шаблонів і символи за спеціальною метрикою. Система знаходить відповідність документів там, де традиційні пошукові системи нічого не знаходить. Тому ця система ідеально підходить для пошуку складних даних, що містяться в «глибинному» Веб.

Також існує багато пошукових систем, що виконують пошук в «глибинному» Веб. Вони спеціалізуються на конкретній предметній області та містять перевірені і рецензовані спеціалістами статті. Такі системи, як правило, мають вузько спрямовані репозиторії, що надає реальну перевагу для цілеспрямованого пошуку дослідника.

До таких спеціалізованих порталів можна віднести Mednar – портал з глибинного пошуку в галузі медицини, Biznar – пошук в бізнес-галузі, Worldwidescience – глобальний науковий портал, Science.gov – науковий портал уряду США, Scitopia – пошукова система наукової інформації і патентів, Nutrition.gov – портал, який містить інформацію про здоров'я. Більшість порталів глибинного вебу підтримують механізми кластеризації за темами.

Висновки

Однією з причин підвищеного інтересу до проекту Semantic Web є практична зацікавленість у поліпшенні якості пошуку у Веб. Дослідження з цієї проблеми ведуться в різних напрямках і дають різноманітні результати у вигляді нових пошукових систем. Такі системи, як Swoogle, дозволяють лише виконувати пошук онтологій за ключовими словами. Але такий сервіс є дуже корисним для розробників семантичних систем і онтологій, хоча він і не розрахований на простого користувача [22]. Джерелами інформації в них служать набори RDF-даних, включаючи дані, пов'язані в рамках проекту Linked Open Data, і мікроформати.

Можна відзначити й інші пошукові системи Semantic Web, багато з яких знаходяться на стадії бета-тестування, тому оцінити їх можливості складно. Деякі системи йдуть по шляху «поглиблення у Веб», інші – більш прискіпливо розвивають алгоритми інтелектуального аналізу та використовують різноманітні джерела інформації про документи, які знаходяться «позадокументом» у Веб. Розвиток технологій інформаційного пошуку призвів до інтенсивного використання мета-інформаційно-пошукових систем; багатоагентних інформаційно-пошукових систем; систем, побудованих на реалізації онтологічних, мовних та управлінських угод і т.п. Більшість пошукових систем йдуть по шляху розвитку персоналізації пошуку, тобто розпізнавання та задоволення потреб користувача.

Традиційні пошукові системи стають все більш точними та об'ємними, однак вони не можуть перевершити інтелект людини. Вони можуть лише порівнювати слова, а не зміст ідеї, яка обговорюються ними. Нові технології пошукових систем 3-го покоління ще знаходяться в стадії формування, але вже зараз вони дають позитивні результати. Нові пошукові системи можуть допомогти зробити пошук більш значущим, суб'єктивним і прив'язаним до задач (task-based), що стоять перед користувачем. Таким чином, розвиток пошукових систем йде по шляху, метою якого є задоволення потреб індивідуального користувача, з його перевагами, характером, рівнем підготовки і знань тощо.

Список літератури

1. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze [Електроний ресурс] / D. Christopher // An

Introduction to Information Retrieval, Online edition (c)2009 Cambridge UP, Draft of April 1, 2009. – Режим доступу до документу: http://www.informationretrieval.org/.

2. Черний Ю.Ю. Школа наукової інформації. Інформаційні потреби. Основи інформаційного пошуку [Електроний ресурс] / Ю.Ю. Черний. – Режим доступу до документу: http://www.bogoslov.ru/text/321597.html.

3. Захаров В.П. Информационно-поисковые системы: учебно-методическое пособие / В.П. Захаров – СПб, 2005. – 320 с.

4. Holscher C. Web search behaviour of Internet experts and Newbies [Електроний ресурс] / C. Holscher, G. Strube // Proceedings of WWW9. – 2000. – Режим доступу до документу: http://www9.org/w9cdrom/81/81.html.

5. Navarro-Prieto R. Cognitive Strategies in Web Searching [Електроний ресурс] / R.. Navarro-Prieto, M Scaife, Y .Rogers // Proceedings of the 5th Conference on Human Factors & the Web. – 1999. – Режим доступу до док.: http://zing.ncsl.nist.gov/hfweb/proceedings/navarro-prieto/index.html.

6. Muramatu J. Transparent queries: Investigating Users' Mental Models of Search Engines / J. Muramatu, W. Prat // Proceedings of SIGIR. – 2001. – P. 121-125.

7. Choo C.W. Information Seeking on the Web – An integrated model of browsing and searching [Електроний ресурс] / C.W .Choo, B. Detlor, D. Turnbull // Proceedings of the Annual Meeting of the American Society for Information Science (ASIS). – 1999. – Режим доступу до документу: http://choo.fis.utoronto.ca/fis/respub/aisis99/.

8. Andrei Broder. A taxonomy of web search [Електроний ресурс] / Andrei Broder // IBM Research. ACM SIGIR Forum archive, Volume 36 , Issue 2 (Fall 2002), ISSN:0163-5840. – P. 3-10. – Режим доступу до документу: http://portal.acm.org/citation.cfm?doid=792550.792552,

9. Cambridge Advanced Learner's Dictionary [Електроний ресурс]. – Режим доступу до документу: http://dictionary.cambridge.org/.

10. Merriam-webster Dictionary. [Електроний ресурс]. – Режим доступу до док.: http://www.merriam-webster.com/.

11. Jack Rodnessy. New Search Engines: The Next Generation of Google Competition [Електроний ресурс] / Jack Rodnessy. – Режим доступу до документу: http://www.webupon.com/Search-Engines/New-Search-Engines-The-Next-Generation-of-Google-Competition.

12. Рогушина Ю.В. Разработка принципов представления электронных изданий, обеспечивающих корректную индексацию поисковыми системами Інтернет / Ю.В. Рогушина, І.Ю. Гришанова // Проблеми програмування. – 2004. – № 4. – С. 39-47.

13. Robin Nobles. The Future Of Search Engine Optimizing: Theme Engines. The next generation of search engines has arrived [Електроний ресурс] / Robin Nobles . – Режим доступу до документу.: http://www.Search-engineworkshops.com/articles/se-optimization-future.html.

14. Андон Ф.И. Semantic web как новая модель информационного пространства Інтернет / Ф.И. Андон, И.Ю. Гришанова, В.А. Резниченко // Проблеми програмування. – 2008. – № 2-3. – С. 417-430.

15. Ebrahim Ezzy. Search 2.0 vs Traditional Search , July 20, 2006. [Електроний ресурс]. – Режим доступу до док.: http://www.readwriteweb.com/archives/search_20_vs_tr.php

16. Shane McLoughlin Searching on the web; the new breed of search engines , May 25, 2009. [Електроний ресурс]. – Режим доступу до документу: http://relativemusings.blogspot.com/2009/05/searching-on-web-new-breed-of-smarter.html.

17. Stephen Wolfram. Wolfram Alpha - computational knowledge engine [Електроний ресурс] / Stephen Wolfram. – 2009. – Режим доступу до документу: http://basetechnology.blogspot.com/2009/03/wolfram-alpha-computational-knowledge.html.

18. Сидоров В. Wolfram Alpha – Computational Knowledge Engine, или Как сложить яблоко с апельсином? [Електроний ресурс] / Валерий Сидоров.. – блог. – 2009. – Режим доступу до док.: http://netler.ru/pc/wolfram.htm.

19. Square your search results with Google Squared. [Електроний ресурс]. – Режим доступу до док.: http://googleblog.blogspot.com/2009/06/square-your-search-results-with-google.html.

20. Сидоров В. Google Squared: как успех Wolfram Alpha взбудоражил Google и что из этого вышло? [Електроний ресурс] / Валерий Сидоров. – блог. – 2009. – Режим доступу до док.:http://netler.ru/pc/google-squared.htm.

21. Summarization, the Answer to Web Search : Interview with Dmitri Soubbotin of SenseBot, Search Engine Journal, December 12th, 2007. [Електроний ресурс]. – Режим доступу до док.: http://www.searchenginejournal.com/ summarization-the-answer-to-web-search-interview-with-dmitri-soubbotin-of-sensebot/6094/

22. Левшин Д.. Web, часть третья [Електроний ресурс] / Дмитрий Левшин // Открытые системы. – 2008. – № 2. – Режим доступу до документу: http://cio.ru/text/print/302/8165094.html

Надійшла до редколегії 1.09.2009

Рецензент: д-р техн. наук, проф. С.В. Козелков, Центральний НДІ навігації і управління, Київ.

РАЗВИТИЕ ТЕХНОЛОГИЙ ИНФОРМАЦИОННОГО ПОИСКА И ПРИМЕНЕНИЕ ИХ В SEMANTIC WEB

И.Ю. Гришанова, С.С. Щербак

Анализируются технологические аспекты внедрения средств информационного поиска в распределенную среду Semantic Web. Предлагаются адаптированные к внедрению в Semantic Web концепции информационного поиска, задачи, модели и классификация систем информационного поиска по разным признакам, варианты внедрения поисковых систем в Semantic Web. Проводится обзор существующих современных поисковых систем, а также предлагается перечень признаков семантических поисковых систем.

Ключевые слова: *Semantic Web, таксономия, информационный поиск, онтологии, поисковые механизмы.*

AN EVOLUTION OF INFORMATION RETRIEVAL TECHNOLOGIES AND ITS APPLICATION IN SEMANTIC WEB

I.J. Grishanova, S.S. Shcherbak

In the paper have analyzed methods and applications of information retrieval in Semantic Web. Basic information retrieval concepts, objectives, models and the classification of information retrieval systems by various features have presented. This paper gives concepts of information retrieval in e new environment of Semantic Web. Examples of search engines that are popular at the moment have considered. In the paper was reduced the attribute list of semantic search engines.

Keywords: *Semantic Web, taxonomy, information retrieval, ontologies, search engine.*