УДК 004.9

Мусієнко М. П., Савінов В. Ю.

РОЗРОБКА СПОСОБУ ПОШУКУ ТА СИСТЕМИ ЙОГО РЕАЛІЗАЦІЇ ДЛЯ РОЗПОДІЛЕНИХ КОМП'ЮТЕРНИХ СИСТЕМ

В статті розроблено спосіб пошуку інформаційних об'єктів, який орієнтований на роботу в розподілених ієрархічних комп'ютерних системах, а також систему реалізації цього способу. Запропоновано використовувати релевантність пошукових слів із використанням рангу запитуючих користувачів, що дозволяє здійснювати коригування пошукових слів. Проведено експериментальну перевірку запропонованої дворівневої пошукової системи.

Ключові слова: *пошукова система, спосіб пошуку, релевантність, ранжування.*

В работе разработан способ поиска информационных объектов, который ориентирован на работу в распределенных иерархических компьютерных системах, а также систему реализации этого способа. Предложено использовать релевантность поисковых слов с использованием ранга запрашивающих пользователей, позволяет осуществлять корректировку поисковых слов. Проведена экспериментальная проверка предложенной двухуровневой поисковой системы.

Ключевые слова: *поисковая система, способ поиска, релевантность, ранжирование.*

In this paper a way of searching information objects was developed, which was designed to work in distributed hierarchical computer systems and system of implementation of this method. Proposed to use the relevance of search words using rank of requesting users, allowing you to make adjustments to the search words. An experimental verification of the proposed two-level search engine.

Key words: *search engine, search method, relevance, ranking.*

Аналіз проблеми і постановка завдання. Одним із основних елементів роботи з інформацією в мережі Інтернет є її пошук. Наприклад, станом на початок 2013 року число документів, проіндексованих тільки пошуковою системою Google, перевищила трильйон документів. Для здійснення пошуку розроблено чимало різноманітних пошукових систем, найбільш вживаними з яких є Yahoo, Google, MSN, Yandex, Rambler і Mail.ru. Однак для пошуку документів, що належать до тієї чи іншої предметної області, користувачі Інтернету нерідко звертаються до тематичних каталогів інтернет-ресурсів – структурованим наборам посилань на документи відповідної тематики. Тобто оновлення систем пошуку, пошукових алгоритмів тощо є вельми актуальною проблемою.

Основними проблемами, з якими стикається користувач, є:

– правильне та чітке формулювання пошукової фрази, що безпосередньо впливає на результат пошуку;

– вибір елементів уточнення, які можуть розширити пошук, або навпаки, звузити;

– велика кількість «інформаційного сміття», що отримується у відповідь на запит.

Як правило, більшість пошукових систем будується за схемою, що показана на рис. 1 [1].

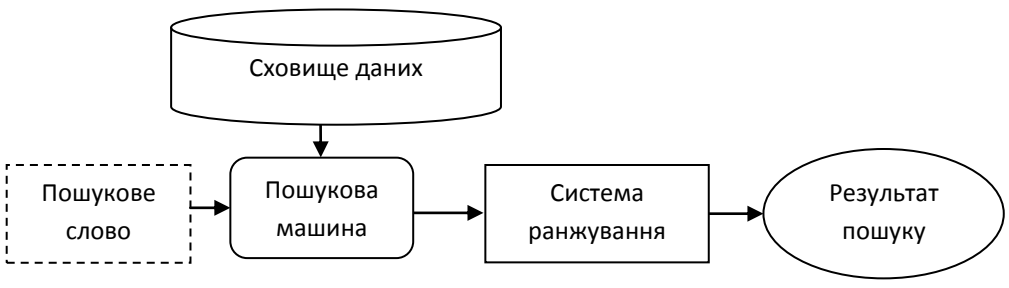

Рис. 1. Схема традиційної пошукової системи

Клієнт формує запит *Q*, який представляється як кортеж слів. Пошукова машина (сервер) отримує запит, робить пошук у своєму банку даних *D* і, виходячи з функції релевантності документа *D* запитом *Q*, ранжує результат, після чого пошукова машина відправляє користувачу відповідь.

Для прикладу на рис. 2 представлено розгорнуту схему функціювання пошукової системи Google.

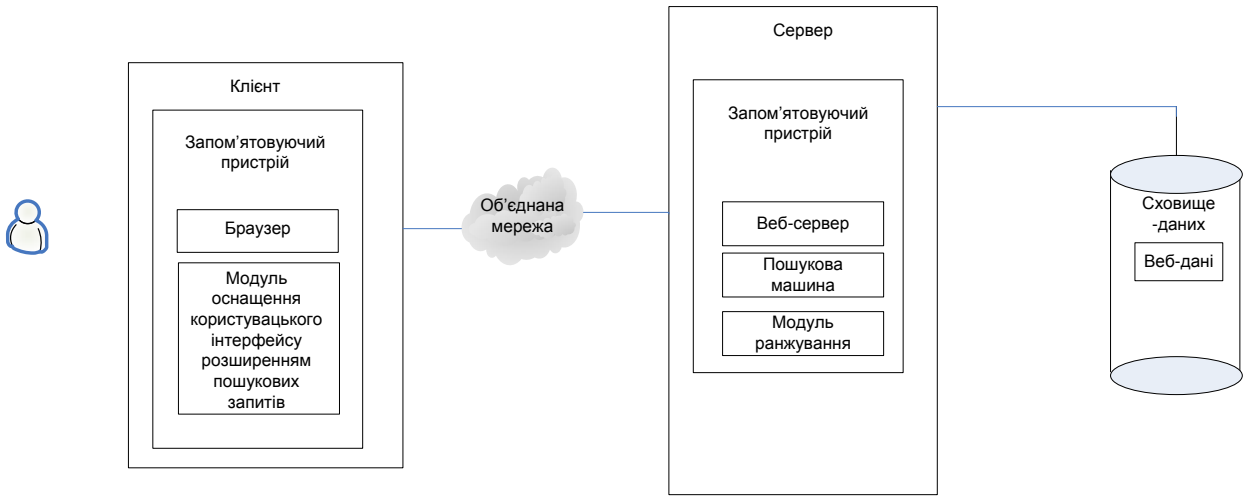

Рис. 2. Блок-схема функціювання пошукової системи Google

Користувачі здійснюють запити зі своїх клієнтських станцій за допомогою браузеру та модулю оснащення користувацького інтерфейсу розширення пошукових запитів, який знаходиться на запам'ятовуючому пристрої. За допомогою об'єднаної мережі запит відправляється серверу, на запам'ятовуючому пристрої якого міститься веб-сервер, з яким браузер клієнта обмінюється веб-інформацією, та який відправляє запит пошуковій машини. Інформація витягаються зі сховища даних, у формі веб-даних, ранжуються за допомогою статистичних показників (наприклад, кількість потрапляння до знайденого документа пошукових слів) та передаються клієнту, що здійснював запит [2].

Як систему ранжирування можна навести найбільш часто використовувану систему Okapi BM25 (базується на TF-IDF функціях ранжирування), що в різних своїх модифікаціях широко застосовується як на Google, Yandex, так і більшості інших пошукових системах. TF-IDF (від англ. TF – term frequency, IDF – inverse document frequency) – статистична міра, яка використовується для оцінки важливості слова в контексті документа. Вага слова пропорційний кількості вживання цього слова в документі, і обернено пропорційна частоті вживання слова в інших документах колекції [2].

Більшість систем (Google, Yandex, тощо) для спрощення процесу пошуку відносно користувача пропонують «живий пошук». Живий пошук – це функція пошуку, яка дозволяє переглядати результати безпосередньо при введенні запиту. Компанією Google було помічено, що введення запиту займає більше часу, ніж прочитання результатів. Між натисканнями клавіш проходить у середньому 300 мілісекунд, а погляд на результати запиту займає всього 30 мілісекунд – у 10 разів менше. Отже, можна одночасно вводити запит і переглядати отримані результати без втрати часу [3].

Головна відмінність «живого пошуку» від звичайного полягає в тому, що користувач отримує потрібну інформацію набагато швидше, так як необов'язково вводити весь запит. Крім того, в «живому пошуку» користувач бачить результати прямо при введенні тексту. Таким чином, користувач може змінювати запит, поки не знайде саме те, що потрібно. Найбільш вагомою перевагою такого пошуку є те, що користувач, не маючи чіткого формулювання ключової фрази, має можливість скористатися накопиченою історією запитів та ключових слів, вибрати з запропонованого списку пошукову фразу [3].

Але існуючі «живі пошуки» мають недоліки. По-перше, такі системи, як Google, Yandex не враховують компетентність користувачів: запропоновані пошукові слова обираються на основі статистичних даних (скільки разів пошукова фраза зустрічалась у запитах), і тому пошукові фрази, які пропонуються користувачеві, не завжди задовольняють його пошуковим намірам. По-друге, до «живого пошуку» включаться ключові слова популярних Інтернет сайтів як пошукові слова, які, як правило, непотрібні користувачу для здійснення пошуку.

Отже, вдосконалення системи пошуку та механізмів для його здійснення, які були б позбавлені зазначених недоліків, є важливою та актуальною задачею.

Метою роботи є вдосконалення способу пошуку інформаційних об'єктів та розробка відповідної пошукової системи, що дозволить покращити релевантність результатів пошуку відносно пошукового запиту користувача і тим самим спростити механізм пошуку.

Викладення основного матеріалу.

У роботі була запропонована ідея використання двох пошукових систем: основної (відомої) і локальної, в якій буде здійснюватися попередня обробка пошукових слів (перед відправкою в основну

пошукову систему) з урахуванням використання накопичених попередніх запитів. При цьому в системі ранжирування локального пошукового модуля використовуватимуться додаткові коефіцієнти, отримані за рахунок попередніх запитів із використанням рангів користувачів, що здійснювали пошук (тобто урахування професійності того, хто робить запит). Це дозволить коригувати поточні запити з урахуванням накопиченого «досвіду» в цій області.

Таким чином, за рахунок коригування запиту в локальному модулі підвищується ймовірність отримання найбільш відповідного запиту пошукового слова. Використання рангу користувачів призведе до того, що вся інформація, що буде накопичуватися в базі даних, буде розподілена не рівномірно, як в існуючих системах, а пріоритетно. Це призводить до підвищення інтелектуалізації системи, тому що з'являється можливість корегувати невірно задані пошукові слова тощо.

Схема такої пошукової системи показана на рис. 3.

Модуль попередньої обробки аналізує запит та вибирає зі сховища даних найближчі запити та впорядковує за рангом користувача, після чого відправляє клієнту обрані запити у вигляді пошукових підказок, для підтвердження запиту. Далі користувач відправляє на сервер остаточний запит, який, у свою чергу, відправляє запит серверу. Необхідна інформація дістається зі сховища даних у формі веб-даних, ранжується за допомогою модуля ранжування і передається веб-серверу та через сервер клієнту, що здійснював запит.

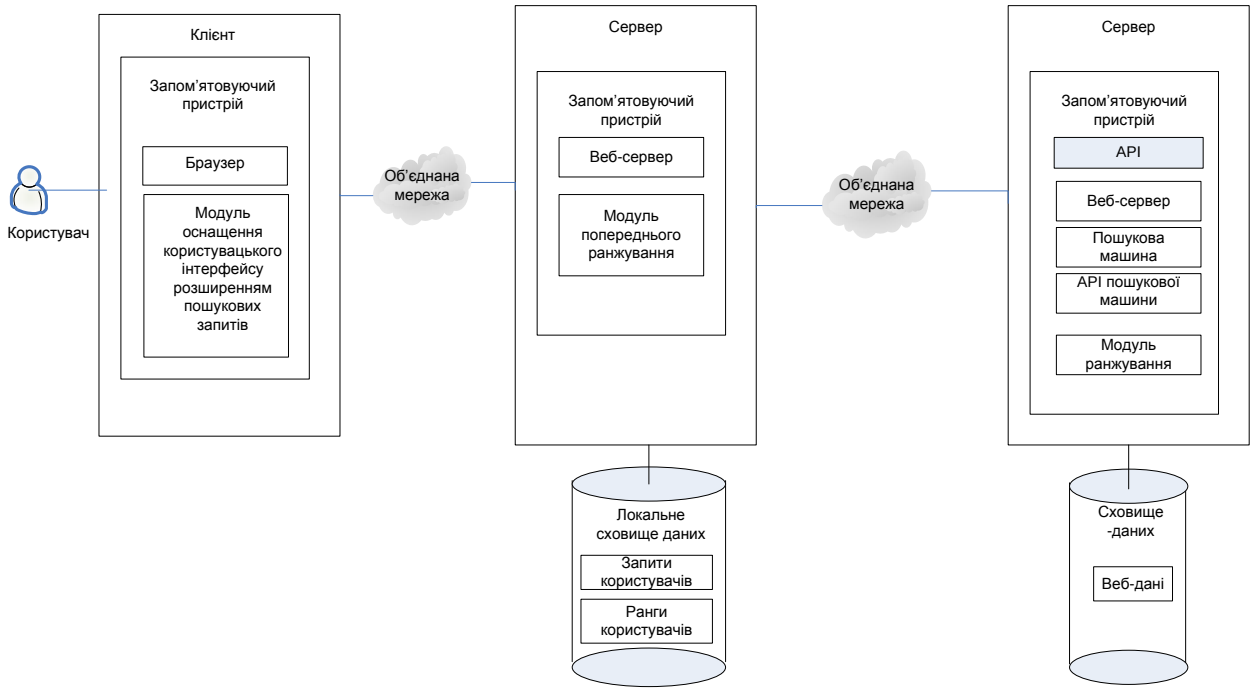

Рис. 3. Схема створеної пошукової системи

Таким чином, користувач формує запит:

$$Q = \{q_1..q_n\}, \quad (1)$$

де $q_1..q_n$ – пошукові слова у пошуковій фразі (запиті), n – кількість слів у запиту.

Загальний список запитів виглядає наступним чином:

$$Q^{all} = \{Q_1^{old}..Q_m^{old}\}, \quad (2)$$

де m – загальна кількість запитів у сховищі даних.

При вводі запиту користувач отримує відранжований список запитів. Кожний такий запит Q_i^{old} має відповідний ваговий коефіцієнт – ранг користувач g_i, де i – порядковий номер запиту. Елемент відранжованого списку визначається таким чином:

$$Q_i^{ranged} = f\{g_i, d(Q, Q_i^{ranged})\}, \quad (3)$$

де $d(Q, Q_i^{ranged})$ – відстань між запитами; i = 1...n, n – кількість накопичених пошукових фраз.

Результатом такої функції буде такий Q_i^{old}, у якого $g_i \to \max$, $d(Q, Q_i^{ranged}) \to \max$.

У свою чергу, g_i, визначається як:

$$g_i = \sum_{j=1}^{n} k_j, \quad (4)$$

де i = 1...n, n – кількість накопичених пошукових фраз; j = 1...m, m – кількість експертних оцінок для визначення компетентності користувача (рангу).

Для перевірки дієздатності запропонованого способу пошуку була розроблена та перевірена пошукова система для обробки спеціалізованої науково-технічної інформації. Була сформована база із запитів викладачів та науково-педагогічних працівників, де як ранги використовувалися такі оцінки, як: науковий ступінь,

кількість наукових публікацій, досвід викладання дисципліни. Як один із прикладів роботи пошукової системи було задано пошукове слово «п'єзодвигун». При використанні відомого способу пошуку інформаційних об'єктів та системи для його здійснення (за рис. 2) пошукова машина видала результат у 107 записів, в якому потрібна інформація (документ) займала місце всередині результуючого масиву інформації. При використанні запропонованого способу пошуку та пошукової системи для його здійснення фраза була відредагована на «ультразвуковий двигун». У результаті пошукова машина на уточнену пошукову фразу видала 67 записів, причому потрібна інформація в результу-ючому масиві була четвертою. Таким чином, при застосуванні запропонованого способу пошукових інформаційних об'єктів отриманий більш високий рівень релевантності пошукових образів знайдених документів до пошукової фрази.

Висновки. У результаті досліджень була розроблена локальна пошукова система, яка може бути використана в розподілених комп'ютерних системах. У сукупності з відомими пошуковими машинами, запропонована система з модулем попереднього ранжування дозволяє проводити корегування запиту, що збільшує результативність пошуку інформації.

Проведені експериментальні дослідження підтвердили працездатність пошукової системи: при розміщенні локального модуля в якості посередника між клієнтом та сервером та при використанні запропонованого «живого пошуку» (з урахуванням рангів користувачів), спрощується процес пошуку інформації, та як наслідок, підвищується релевантність знайдених документів.

ЛІТЕРАТУРА

1. Мусиенко М. П. Разработка дистанционного управляемых робототехнических систем с локальным поисковыми модулями / М. П. Мусиенко, В. Ю. Савинов, А. В. Россоха // Вісник Черкаського державного технологічного університету. – 2012. – № 4. – С. 35–38.
2. Патент 90764. Україна, МПК(2009) G06F17/30, G06F7/00, G06F12/00. Спосіб пошуку інформаційних об'єктів та система для його здійснення / Вакарін Сергій Ігоревич, Небелиця Віталій Миколайович. – № а200806361; заявл. 13.05.2008; опубл. 25.11.2009, бюл. № 22.
3. Google inc. [Электронний ресурс] : Google О Живом поиске. – Режим доступа : http://www.google.ru/instant/.

Рецензенти: Кондратенко Ю. П., д.т.н., професор;
Гожий О. П., к.т.н., доцент.

Дата надходження статті до редколегії 10.05.2013 р.

МУСІЄНКО Максим Павлович, д.т.н., професор, завідувач кафедри інформаційних технологій Чорноморського державного університету імені Петра Могили.

САВІНОВ В. Ю., аспірант кафедри інформаційних технологій Чорноморського державного університету імені Петра Могили.