УДК 004.415, 004.738.5

Н.О. Думанський, О.В. Марковець
Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж

ІНТЕЛЕКТУАЛЬНА СИСТЕМА ПОШУКУ ТА ЗБИРАННЯ ІНФОРМАЦІЇ З ТЕМАТИЧНИХ ВЕБ-РЕСУРСІВ

Розглянуто та проаналізовано загальні принципи пошуку інформації. Розроблено та описано алгоритми функціонування системи пошуку та збирання інформації з тематичних веб-ресурсів.

In this article is conducted general principles of information retrieval and their analysis. Developed and described algorithms of functioning of the system of search and collection of information from thematic web-resources.

Вступ

У сучасному світі Інтернет вже давно використовують як широкий довідковий інструмент. За останні роки він став середовищем опрацювання та зберігання наукової, бізнесової та інших типів інформації. Але основними особливостями Інтернету є динаміка інформації, її постійне оновлення та поширення по всьому Інтернету.

Саме тому пошукові системи вже давно стали невід'ємною частиною Інтернету. Завдяки ним користувачі Всесвітньої павутини намагаються знайти потрібну інформацію. Переважно, пошук інформації зводиться до пошуку сторінки, на якій розміщена ця інформація, і таких сторінок в Інтернеті може бути декілька. Для отримання результатів, які дадуть змогу порівняти інформацію на однотипних веб-ресурсах, потрібно створити тематичні пошукові системи.

Постановка проблеми та аналіз останніх досліджень

Створення тематичної пошукової системи містить в своїй основі такі задачі:

- підбір тематичних сайтів,
- пошук інформації на різних ресурсах з різними запитами,
- видобування інформації з сайту,
- створення єдиної системи представлення інформації користувачу.

Враховуючи поставлені задачі, алгоритм роботи тематичного пошуковика схожий з алгоритмом звичайної пошукової системи.

Пошукова система складається з таких основних компонентів:

- Spider (павук) – браузероподібна програма, яка викачує веб-сторінки.
- Crawler (краулер, «мандрівний» павук) – програма, яка автоматично проходить по всіх посиланнях, знайдених на сторінці.
- Indexer (індексатор) – програма, яка аналізує веб-сторінки, викачані павуками.
- Database (база даних) – сховище викачаних та опрацьованих сторінок.
- Search engine results engine (система видачі результатів) – витягує результати пошуку з бази даних.
- Web server (веб-сервер) – веб-сервер, який здійснює взаємодію між користувачем та іншими компонентами пошукової системи.

Web server та Search engine results engine часто називають просто пошуковим сервером.

У деталях реалізації пошукових механізмів можуть відрізнятися одна від однієї (наприклад, зв’язка Spider+Crawler+Indexer може бути виконана у вигляді єдиної програми, яка викачує відомі веб-сторінки, аналізує їх і шукає за посиланнями нові ресурси), проте всім пошуковим системам властиві описані загальні риси.

Spider. Забезпечує скачування сторінки і витягує всі внутрішні посилання з цієї сторінки. Викачується html-код кожної сторінки. Для скачування сторінок роботи використовують протоколи HTTP. Працює «павук» так. Робот передає на сервер запит “get/path/document” і деякі інші команди HTTP-запиту. У відповідь робот отримує текстовий потік, що містить службову інформацію і безпосередньо сам документ.

Посилання витягуються з тегів a, area, base, frame, frameset та ін. Крім посилань, роботи обробляють редиректи (перенапрямлення). Кожна викачана сторінка зберігається у такому форматі:

- URL сторінки
- дата, коли сторінка була викачана
- http-заголовок відповіді сервера
- тіло сторінки (html-код)

Crawler. Виділяє всі посилання, присутні на сторінці. Його завдання – визначити, куди далі повинен йти павук, ґрунтуючись на посиланнях або за заздалегідь заданим списком адрес. Краулер, проходячи по знайдених посиланнях, шукає нові документи, ще не відомі пошукові системи.

Indexer. Індексатор розбирає сторінку на складові частини і аналізує їх, застосовуючи власні лексичні і морфологічні алгоритми. Аналізу піддаються різні елементи сторінки, такі як текст, заголовки, посилання структурні і стильові особливості, спеціальні службові html-теги і так далі.

Отже, модуль індексування дає змогу обходити за посиланнями задану кількість ресурсів, викачувати сторінки, що зустрічаються, витягувати посилання на нові сторінки з отриманих документів і здійснювати повний аналіз цих документів.

Database. База даних – це сховище всіх даних, які пошукова система викачує і аналізує. Інколи базу даних називають індексом пошукової системи.

Пошуковий сервер є найважливішим елементом всієї системи, оскільки від алгоритмів, покладених в основу її функціонування, безпосередньо залежить якість і швидкість пошуку.

Пошуковий сервер працює так:

- Отриманий від користувача запит піддається морфологічному аналізу. Генерується інформаційне оточення кожного документа, що міститься в базі (яке і буде згодом відображено у вигляді відповідної текстової інформації на сторінці видачі результатів пошуку).
- Отримані дані передаються як вхідні параметри спеціальному модулю ранжирування. Обробляються дані за всіма документами, внаслідок чого для кожного документа розраховується власний рейтинг, що характеризує релевантність запиту, введеного користувачем, і різних складових цього документа, що зберігаються в індексі пошукової системи.
- Залежно від вибору користувача цей рейтинг може бути скоригований додатковими умовами (наприклад, так званий «розширений пошук»).
- Далі генерується сніппет, тобто для кожного знайденого документа з таблиці документів витягуються заголовок, коротка анотація, яка найбільше відповіднає запиту і посилання на сам документ, причому знайдені слова підсвічують.
- Отримані результати пошуку передаються користувачеві у вигляді сторінки видачі пошукових результатів.

Основні характеристики пошукових систем

- Точність – ще одна основна характеристика пошукової системи, яка визначається рівнем відповідності знайдених документів запиту користувача. Наприклад, якщо за запитом «як вибрати автомобіль» знаходять 100 документів, в 50 з яких є словосполучення «як вибрати автомобіль», а в інших лише наявність цих слів («як правильно вибрати автомагнітолу та встановити її в автомобіль»), то точність пошуку буде 50/100 (=0,5). Чим точніший пошук, тим швидше користувач знайде потрібні йому документи, тим менше різного роду «сміття» буде в них зустрічатись, тим рідше знайдені документи не відповідатимуть запиту.

- Актуальність – не менш важлива складова пошуку, яка характеризується часом, що пройшов від моменту публікації документа в мережі Інтернет, до занесення його до індексної бази пошукової системи. Наприклад, на наступний день після появи цікавої новини, велика кількість користувачів звернулись до пошукових систем з відповідним запитом. Об'єктивно з моменту публікації новини на цю тему пройшло менше доби, однак основні документи вже були проіндексовані та доступні для пошуку завдяки існуванню у великих пошукових системах так званої «швидкої бази», яка поновлюється декілька разів на день.

Загальна модель функціонування тематичної пошукової системи

Принцип роботи розробленої тематичної пошукової системи наведено на рис. 1.

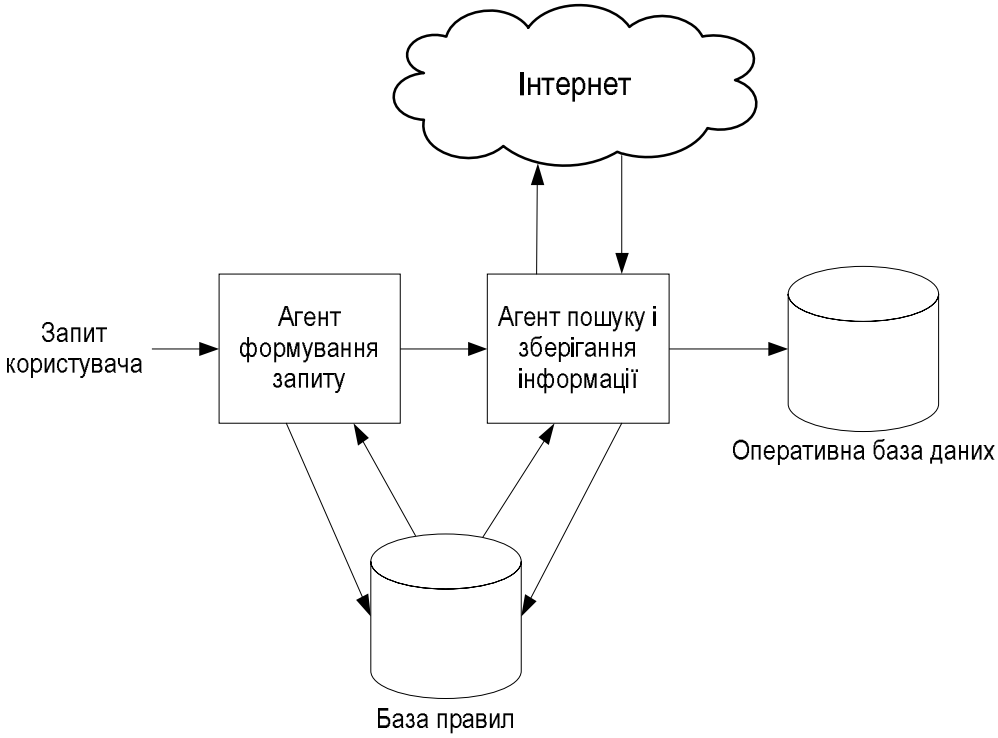

Рис. 1. Схема роботи тематичної пошукової системи

Користувач, використовуючи можливості клієнтської частини, вводить ключове слово для пошуку інформації та вибирає із списку тематичних сайтів ті, з яких потрібно отримати інформацію за ключовим словом.

Агент формування запиту, отримавши ключове слово і список вибраних сайтів, звертається до бази правил для отримання правил формування запитів до вибраних сайтів. Отримавши ці правила, формує запит до кожного з вибраних сайтів і цю інформацію передає агенту пошуку і зберігання інформації. Своєю чергою, агент пошуку та зберігання інформації відсилає отримані запити в Інтернет і отримує html-коди сторінок, які записує в свою оперативну базу даних. Потім цей агент запитує в бази правил правила видобування інформації з html-кодів. Опрацьовує інформацію, записану в своїй оперативній базі, згідно з правилами і записує вибрану за правилами інформацію в оперативній базі даних системи. Далі з цією базою даних працює клієнтська частина.

Що ж являють собою ці бази даних? База правил формується розробником системи. Саме розробник опікується її правильним наповненням. Сама база правил складається з таких полів:

- назва сайту (єдине, що відображається користувачеві в клієнтській частині);
- приклад формування запиту – тут наведено приклад запиту до відповідного сайту;
- інформація про правило поєднання ключового слова із запитом попереднього поля;

- правила видобування інформації з html-коду отриманої в результаті пошуку сторінки відповідного сайту.

Враховуючи динаміку руху інформації в Інтернеті, база правил повинна постійно перебувати під контролем розробника для зазначення відповідних змін властивостей сайту, для внесення нових тематичних сайтів, що з'явилися в мережі. Для наповнення цієї бази можна також використовувати інтелектуальні системи, які б в кооперації зі стандартними системами пошуку відбирали сайти з конкретною тематикою. Але часто для розкрутки своїх сайтів веб-майстри використовують не дуже чесні прийоми, які можуть давати неправильні результати аналізу веб-ресурсу. Саме для цього у процесі наповнення бази правил використовують людський фактор, за допомогою якого коректно відсіюють «непридатні» сайти.

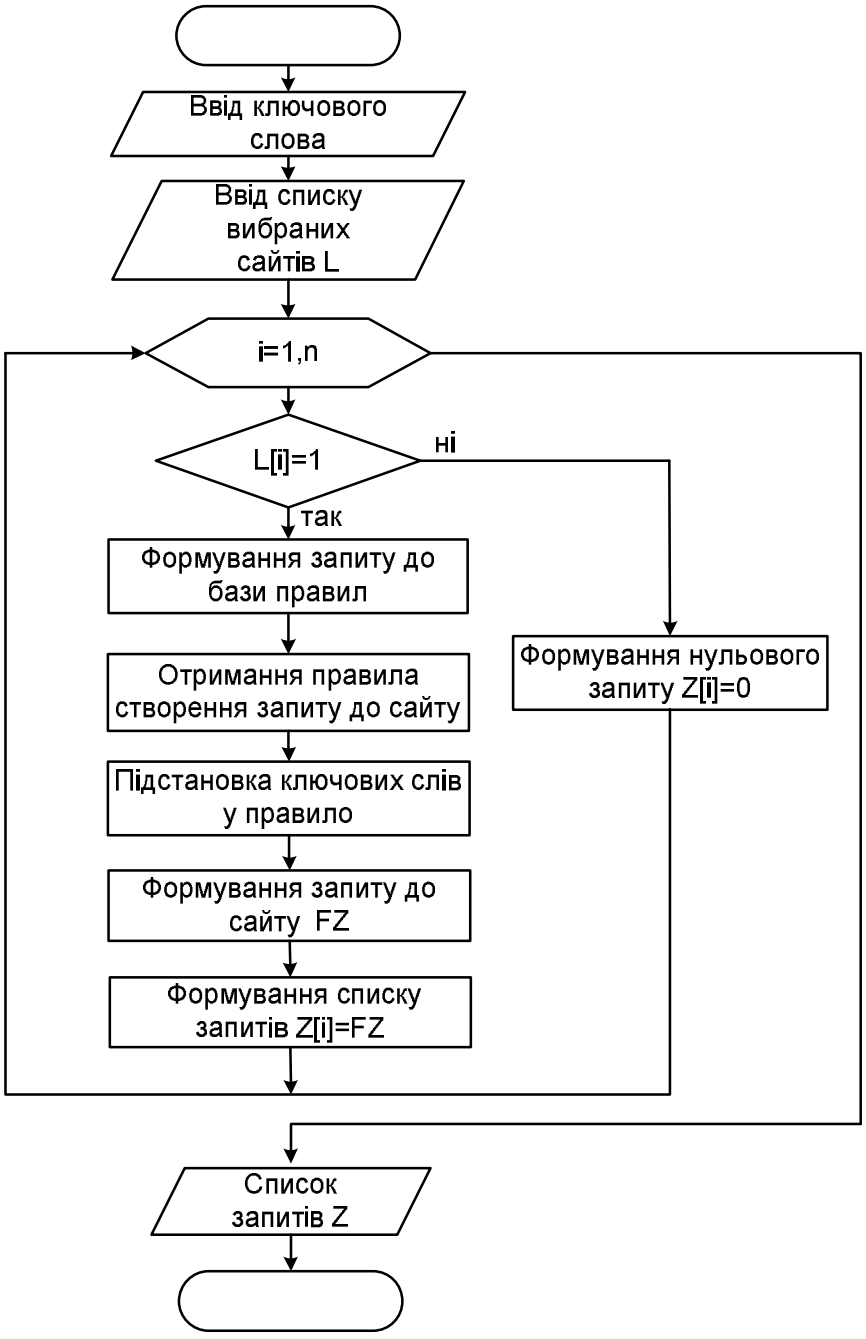

Рис. 2. Блок-схема алгоритму роботи агента формування запиту

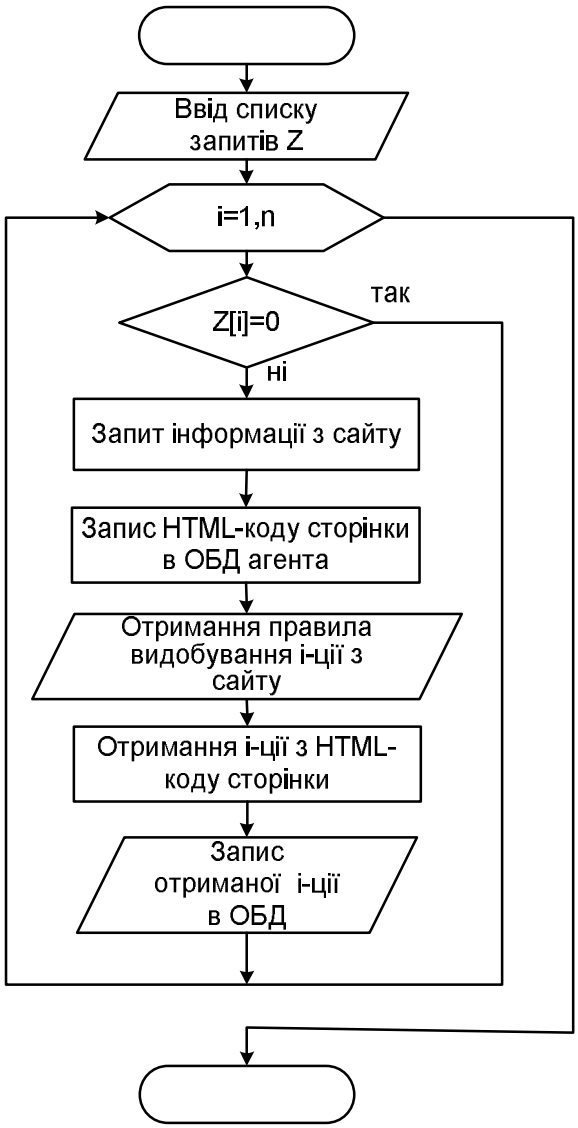

Рис. 3. Блок-схема алгоритму роботи агента пошуку і зберігання інформації

Оперативна база даних складається з інформації, необхідної користувачу, яка є невпорядкованою. Агент пошуку і зберігання інформації записує в оперативну базу даних отриману з html-коду інформацію, яка зберігається під такими полями:

- назва сайту;
- зображення, що відповідає конкретному результату;
- текст результату пошуку;
- посилання на детальнішу інформацію;

Ця база даних формується для оперативного збереження даних, які користувач може переглядати, сортувати, відбирати за певними критеріями, використовуючи клієнтську частину системи.

Розглянемо алгоритм роботи кожного агента окремо.

Агент формування запиту (рис. 2):

1. Отримання ключового слова від клієнтської частини.
2. Зчитування списку зазначених користувачем сайтів пошуку.
3. Перевірка кожного елементу зі списку сайтів (позначений – крок 4, позначений – крок 9).
4. Формування запиту до бази правил щодо вибраного сайту.
5. Отримання правила формування запиту до конкретного сайту.

6. Внесення ключового слова пошуку у правило формування запиту.
7. Створення запиту до вибраного сайту.
8. Формування списку запитів для агента пошуку та зберігання інформації (перехід на крок 10).
9. Додавання до списку запитів нульового запиту.
10. Повернення на крок 3, поки не закінчиться список сайтів.
11. Передача списку запитів агенту пошуку та зберігання інформації.

Агент пошуку і зберігання інформації (рис.3):

1. На початку своєї роботи отримує від внутрішнього агента список сформованих запитів до тематичних сайтів.
2. Перевірка, чи запит не нульовий (якщо так – крок 8).
3. Відсилання запиту тематичному сайту.
4. Одержання коду html-коду сторінки, яка сформувалась як відповідь на запит.
5. Звернення до бази правил та отримання правила видобування інформації з коду сторінки.
6. Отримання потрібної інформації з html-коду сторінки (малюнки, текст, посилання).
7. Запис отриманої інформації в оперативну базу даних системи.
8. Перехід на крок 2 до закінчення списку запитів.

Як видно, всі ці компоненти тісно пов'язані один з одним і працюють у взаємодії, утворюючи чіткий, достатньо складний механізм роботи тематичної пошукової системи, що вимагає величезних витрат ресурсів.

Висновок

Відмінність нашої системи від загальноприйнятих систем пошуку полягає в тому, що вона не відображає користувачу посилання на сторінки, які його цікавлять, а відображає саму інформацію з цих сторінок, оскільки тематика пошуку нам вже відома, а це дає змогу звузити коло до лише достовірних результатів.

Така система тематичного пошуку може використовуватись для вирішення багатьох задач:

- виявлення оптимальної пропозиції купівлі/продажу з кількох Інтернет магазинів або аукціонів;
- наповнення методичного матеріалу різноманітних курсів дистанційного навчання з Інтернет енциклопедій та довідників;
- пошук бізнес-партнерів у різних галузях;
- порівняння пропозицій роботодавців з різних мережевих дошок оголошень для молодих спеціалістів та багатьох інших задач, які у своїй основі вимагають малої затрати часу, оптимальних рішень та наочності результатів.

1. Гладун А.Я. Онтологии как перспективное направление интеллектуализации поиска информации в мультиагентных системах е-коммерции / А.Я. Гладун, Ю.В. Рогушина // Proc. of XI-th International Conference «Knowledge-Dialogue-Solution». – Vol. 1. – Varna, 2005. 2. Andon Ph., Deretsky V. Control Oriented Ontology and Process Description for Cooperation Agents in Information Retrieval // Sixth International Scientific Conference „Electronic Computers and Informatics ECI'2004". − Kosice – Herlany, Slovakia; September 22-24, 2004. 1.Google's Secret Lab June 6, 2005 доступне по посиланню http://www.markcarey.com/ googleguy-says/archives/googles-secret-lab.html 3. List of Google products from Wikipedia, the free encyclopedia доступне по посиланню http://en.wikipedia.org/wiki/List_of_Google_products.