УДК 004.04

СКАНУВАННЯ АГРЕГОВАНИХ ДАНИХ

Я.С. Колєснікова[1], *Г.Е. Заволодько*[2]

[1] *магістрант кафедри Системи інформації ім. В.О. Кравця, НТУ «ХПІ», Харків, Україна*

[2] *доцент кафедри Системи інформації ім. В.О. Кравця, канд. техн. наук, НТУ «ХПІ», Харків, Україна*

anna.zavolodko@khpi.edu.ua

Веб-сканер — це інтернет-бот, також відомий як веб-павук, автоматичний індексатор або веб-робот, який працює для систематичного сканування мережі. Ці боти майже як архівісти та бібліотекарі Інтернету.

Вони об'єднують і завантажують інформацію та вміст, який потім індексується та каталогізується в результатах пошуку, щоб користувачі могли бачити його в порядку відповідності. Ось як пошукова система, така як Google, може швидко відповідати на пошукові запити користувачів саме тим, що ми шукаємо: застосовуючи свій алгоритм пошуку до даних веб-сканера [1].

Щоб знайти найбільш достовірну та релевантну інформацію, бот починає роботу з певного вибору веб-сторінок. Він шукатиме (або сканує) ці дані, а потім переходитиме за посиланнями, згаданими в них (або павуком), на інші сторінки, де повторюватиме те ж саме. Зрештою сканери створюють сотні тисяч сторінок, інформація з яких потенційно може відповісти на ваш пошуковий запит.

Наступним кроком для таких пошукових систем, як Google, є ранжування всіх сторінок за певними факторами, щоб надати користувачам лише найкращий, найнадійніший, найточніший і найцікавіший вміст. Факторів, що впливають на алгоритм Google і процес ранжування, багато і вони постійно змінюються. Деякі з них більш відомі (ключові слова, розміщення ключових слів, внутрішня структура посилань і зовнішні посилання тощо). Інші складніше визначити, як, наприклад, загальну якість веб-сайту.

Деякі сканери також створюють графічне представлення завантажених сайтів (графік, де вузли є сторінками, а краї представляють гіперпосилання).

Поведінка веб-сканера є результатом комбінації політик [2]:

Політика вибору, яка визначає, які сторінки завантажувати. Якщо набір S реалізовано як стек, алгоритм відвідує веб-сайт у глибину, тоді як використання черги забезпечує відвідування в ширину.

Політика ввічливості. Використання веб-сканерів корисно для багатьох завдань, але має певну ціну для загальної спільноти. Витрати на використання веб-сканерів включають мережеві ресурси та перевантаження серверів. Частковим вирішенням цих проблем є файл виключення robots. Цей файл вказує, до яких веб-сайтів (або їх частин) сканери не повинні отримати доступ.

Політика розпаралелювання, яка визначає, як координувати процеси або потоки на етапі завантаження.

Веб-сканер повинен мати хорошу стратегію сканування, але він також потребує оптимізованої архітектури. Шкапенюк і Суел зазначили, що: «Хоча досить легко побудувати повільний сканер, який завантажує кілька сторінок на секунду протягом короткого періоду часу, побудувавши високопродуктивну систему, яка може завантажувати сотні мільйонів сторінок протягом кількох тижнів представляє низку

проблем у дизайні системи, ефективності введення/виведення та мережі, а також надійності та керованості».

Як згадувалося вище, основною функцією веб-сканера є додавання нових URL-адрес до межі та вибір наступної URL-адреси з цієї межі для подальшої обробки після кожного рекурсивного кроку. Взагалі існує кілька методів сканування, які використовуються в роботі веб-сканерів.

Сканування загального призначення Використовуючи цей метод сканування, веб-сканер збирає якомога більше сторінок із заданого набору URL-адрес і гіперпосилань із них. Таким чином сканер може отримати велику кількість сторінок з різних «місць» Інтернету. Однак варто зазначити, що сканування загального призначення може значно сповільнити швидкість і пропускну здатність мережі. Це тому, що всі сторінки витягуються за допомогою цього методу сканування.

Сфокусоване сканування Основна відмінність між сфокусованим скануванням і скануванням загального призначення полягає в тому, що перше призначене для збору сторінок лише на певну тему, яка визначається перед запуском веб-сканера. Ця функція економить фізичні та мережеві ресурси.

Розподілене сканування Цей метод сканування дозволяє використовувати кілька процесів для сканування та індексування сторінок з Інтернету. Розподілене сканування призводить до невеликої економії пропускної здатності мережі. Однак використання кількох комп'ютерів може уникнути витрат, які інакше були б використані для підтримки обчислювальних кластерів [3].

Повноту результатів сканування можна визначити як кількість сторінок, завантажених сканером, із загальної кількості веб-сторінок, що складають веб-додаток (тобто збережених на веб-сервері або згенерованих програмою веб-сервера). Оскільки останнє число може бути невідомим, практичний спосіб обчислення відносної повноти веб-сканера для порівняння з набором інших сканерів полягає в тому, щоб взяти лише загальну кількість сторінок, завантажених даним сканером. Це можна порівняти із загальною кількістю сторінок, завантажених іншими сканерами під час запуску на тому самому сайті. Повнота дуже важлива, оскільки низький рівень повноти може призвести до «часткового» завантаження веб-програми, погіршуючи продуктивність сканера у таких важливих завданнях, як: побудова структурної моделі чи індексу, віддзеркалення, тощо [4].

Пошук усіх гіперпосилань на веб-сторінці та їх автоматичне вирішення є нетривіальним завданням. Це тому, що підтримувані на даний момент мови (HTML, Javascript) дозволяють різні способи (синтаксичні варіанти) вставки гіперпосилання на веб-сторінку.

Надійність можна визначити як здатність сканера функціонувати правильно (тобто без збоїв) за наявності: великих веб-додатків; синтаксично недійсних гіперпосилань; проблеми аналізу/сканування.

Список літератури:

1. *Заволодько, Г. Е.; Касілов, О. В.* Інтерактивні засоби в онлайн-освіті. Цифрова платформа: інформаційні технології в соціокультурній сфері, 2020, 3,№ 1: 11-21

2. *Singh, Apoorv Vikram; Vikas, Achyut Mishra.* A review of web crawler algorithms. International Journal of Computer Science & Information Technologies, 2014, 5.5: 6689-6691.

3. *Kausar, Md Abu; Dhaka, V. S.; Singh, Sanjeev Kumar.* Web crawler based on mobile agent and java aglets. IJ Information Technology and Computer Science, 2013, 5.10: 85-91.

4. *Заволодько Г.* агрегатор онлайн-курсів в навчальному процесі / *Заволодько Ганна, Королех Єлизавета* // Project approach in the didactic process of universities - international dimension. - 2021. - № III(V). - с. 271 – 283