УДК 681.3.06

ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ ПОШУКУ ДОКУМЕНТІВ В ІНТЕРНЕТ ІЗ ВРАХУВАННЯМ ПОДІБНОСТІ ВЕБ-СТОРІНОК

Мельник А.М.[1)], Миць О.З.[2)]
Тернопільський національний економічний університет
[1)] к.т.н., доцент; [2)] магістрант

I. Постановка проблеми

Велика кількість даних та технологій «розкрутки» сайтів призводить до необхідності розробки методів та алгоритмів фільтрації результатів пошуку з метою покращення ефективності пошуку в Інтернеті [1]. Пошукові системи використовують різноманітні коефіцієнти (Google – PR, Яндекс – ТІЦ) для сортування результатів пошуку. Дуже часто частина сайтів містить ідентичну інформацію, що призводить до втрачання часу користувачами на фільтрування результатів пошуку [1].

II. Мета роботи

Метою дослідження є розробка методу оцінки подібності веб-сторінок та його практична реалізація в системі інформаційного пошуку.

III. Врахування подібності веб-сторінок в системах інформаційного пошуку

Під подібністю веб-сторінок будемо розуміти ступінь дублювання даних, які вони містять

$$S(P_1, P_2) = D_1 \cup D_2, \tag{1}$$

де P_1, P_2 - веб-сторінки, D_1, D_2 - контент відповідних сторінок.

Нехай f_1, f_2 - набір унікальних ознак сторінок P_1, P_2. Під унікальною ознакою будемо розуміти набір ознак, що не повторюються в межах однієї веб-сторінки. В якості таких ознак будемо використовувати елементи інформаційного наповнення, а саме слова, речення, посилання та графічний контент. Загальний набір ознак F утворюється шляхом об'єднання набору ознак сторінок, що порівнюються між собою. Ознаки, що повторюються, також видаляються з набору.

$$F = f_1 \cup f_2, count(F) \le \left(count(f_1) \cup count(f_2)\right), \tag{2}$$

Коефіцієнт дублювання даних i-ї веб-сторінки в j-й обчислюємо наступним чином

$$\delta_{P_i(P_j)} = \frac{count(f_i) + count(f_j) - count(F)}{count(f_i)}. \tag{3}$$

Загальний коефіцієнт даних, що дублюються в веб-сторінках P_1, P_2 обчислюємо за допомогою виразу

$$\delta_{P_i,P_j} = \frac{count(f_i) + count(f_j) - count(F)}{count(f_i) + count(f_j)}, \tag{4}$$

а коефіцієнт подібності відповідно як

$$S_{P_i,P_j} = 2 \times \delta_{P_i,P_j}, S_{i,j} \in [0,1]. \tag{5}$$

Відповідно відстань між веб-сторінками можна обчислити наступним чином

$$D_{P_i,P_j} = \frac{1}{S_{P_i,P_j}}. \tag{6}$$

Таким чином, коефіцієнт подібності враховує невідповідність розмірів веб-сторінок та їх величини, а відстань між сторінками є певним чином обмеженою, так як коефіцієнт подібності може змінюватися лише в діапазоні від 0 до 1, що дозволяє вирішити проблему розмірності відстані та подібності.

Висновок

У роботі досліджено вплив подібності веб-сторінок на швидкість опрацювання результатів інформаційного пошуку.

Список використаних джерел

1. Краковецький О. Ю. Метод оцінки подібності веб-сторінок / В. М. Дубовой, О. Ю. Краковецький, О. В. Глонь // Оптико-електронні інформаційно-енергетичні технології. – 2008. – №2 (16). – С. 5-9.