УДК 044.03; 681.518:061

А.В. Порошин

Харківський національний університет радіоелектроніки, Харків

ДОСЛІДЖЕННЯ МОДЕЛЕЙ РАНЖУВАННЯ ДОКУМЕНТІВ У МЕТАПОШУКОВИХ СИСТЕМАХ

Стрімкий розвиток та розбудова глобальної комп'ютерної мережі Інтернет створили принципово нові умови та можливості застосування інформаційних технологій для доступу, пошуку, опрацювання та зберігання інформації.

За таких обставин для ефективного пошуку потрібної (релевантної) інформації необхідні автоматизовані інформаційно-пошукові системи, які ґрунтуються на інтелектуальних алгоритмах аналізу текстів.

Аналіз існуючих підходів до автоматизації інформаційного пошуку текстових документів та напрямків розвитку провідних пошукових систем світу, засвідчив переваги метапошукових систем (МПС) в сфері пошуку релевантної інформації [1].

Формалізована модель МПС має вигляд:

$$M_{МПС} = \{M_{МПМ}, Z, I\},$$

де $M_{МПМ}$ – модель метапошукової машини, Z – модель засобів перегляду документів, I – модель інтерфейсу.

Модель метапошукової машини має вигляд

$$M_{МПМ} = \{V, Q, F, R(\upsilon_i, q_i)\},$$

де V – множина документів, що відомі усім інформаційно-пошуковим системам (ІПС), у базі яких здійснюється пошук, Q – множина пошукових запитів, які після введення трансформуються у відповідності до вимог кожної окремої ІПС, $F = \{\cup, \cap, /\}$, де $\cup$ – об'єднання, $\cap$ – перетин, $/$ – різниця, а $R(\upsilon_i, q_i)$ – функція ранжування.

Така модель МПС дозволяє виділити основні переваги та недоліки МПС в цілому. Основною перевагою МПС за цією моделлю є можливість пошуку інформації одночасно в декількох пошукових системах, каталогах та, іноді, у «невидимій павутині» (збірці онлайнової інформації, яка не проіндексована звичайними пошуковими системами). Основними недоліками МПС за цією моделлю, які значно знижують потенціал їх використання, є проблема невідповідності пошукових мов різних пошукових систем одна іншій та проблема підвищення якості формування моделей ранжування документів.

У загальному випадку, функція ранжування пошукової системи для Інтернету має такий вигляд:

$$Rel_{\omega}(\tau) = P(\tau, \omega) \cdot q(\tau),$$

де $P(\tau, \omega)$ – показник документу τ для запиту ω по внутрішніх чинниках, тобто показник релевантності документу τ по відношенню до запиту ω, а $q(\tau)$ – якісна функція від τ, яка розраховується на підставі зовнішніх чинників, не представлених безпосередньо на самій сторінці [2].

Оскільки показник $P(\tau, \omega)$ залежить лише від внутрішніх чинників, які можуть зазнавати людського впливу, якість ранжування на основі лише цього показника не може бути задовільною, що і стало основною причиною введення у функцію ранжування показника $q(\tau)$, який дозволяє оцінювати якість сторінки незалежно від запиту.

А тому саме від правильності вибору функції $q(\tau)$ і буде залежати якість ранжування документів та адекватність представленої користувачеві вибірки документів.

Припускаючи, що користувач, виконуючи запит до пошукової системи, жадає знайти авторитетну інформацію раніше за все інше, для розрахунку показника $q(\tau)$ пропонується використовувати показник авторитетності Page Rank (PR), який дозволяє оцінювати якість сторінки за кількістю та якістю ресурсів, що на неї посилаються.

Припускається, що модель вигляду

$$Rel_{\omega}(\tau) = P(\tau, \omega) \cdot q(\tau, PR)$$

дозволить підвищити якість ранжування при формуванні вихідної вибірки документів.

Список літератури

1. Иванов, В.К. Введение в информационно-поисковые системы [Текст] / В.К. Иванов. – Тверь: ТГТУ, 2005 – 37 с.

2. Braslavski, P. Meta-search, thesaurus and GUI for focused WEB information retrivial [Текст] / P. Braslavski. – Пушкино: Digital Libraries, 2004 – 140 с.