УДК 022.4.001.57

Ю. В. Яковлєва
Національна бібліотека України ім. В.І.Вернадського
проспект 40-річчя Жовтня, 3, 03039 Київ, Україна

Методика ранжування результатів пошуку в інформаційно-пошукових системах бібліотек

Запропоновано методику ранжування результатів пошуку в інформаційно-пошукових системах науково-технічних бібліотек, що ґрунтується на багатокритеріальній оцінці видань за їх інформативністю.

Ключові слова: *інформаційно-пошукові системи, статистичний аналіз, автоматизація бібліотечних процесів.*

Вступ

У сфері інтелектуалізації інформаційних систем велика увага приділяється інтелектуальним технологіям ідентифікації (нечіткі множини, генетичні алгоритми, нейронні мережі) [1, 2], оцінюванню і розпізнаванню зображень [3], розробці методів представлення й обробки некоректних даних і знань [4, 5] та іншим напрямкам. Значну увагу теоретичних досліджень і практичних розробок спрямовано на побудову інтелектуальних адаптивних користувацьких інтерфейсів [6–8]. Питанню інтелектуалізації бібліотечних інформаційно-пошукових систем (ІПС) належної уваги не приділяється. Одним із важливих напрямів інтелектуалізації ІПС бібліотек, поряд з розвитком лінгвістичного забезпечення й організацією інтелектуального природномовного користувацького інтерфейсу, є організація зворотного зв'язку в ІПС на основі використання статистичних даних для оцінки релевантності результатів пошуку. Метою даної статті є розробка методики ранжування результатів пошуку в інформаційно-пошукових системах науково-технічних бібліотек

Ранжування результатів інформаційного пошуку

Доцільність використання і розвитку найцінніших традиційних технологій, зокрема бібліотечних, постійно підтверджується онлайновими інформаційно-пошуковими системами, які мають унікальні можливості випробовування нових методик і швидкого реагування на зовнішні впливи, оскільки необмежена повнота

пошуку не стримує спроб збільшення релевантності за рахунок підвищення точності пошуку і використання інших технологій. Так, відомий в бібліотекознавстві метод контент-аналізу, який є одним із методів якісно-кількісного вивчення змісту текстів, послугував аналогом методу пошуку Web-сторінок на основі аналізу метаданих, розміщених у заголовках HTML-документів. Так само знайшли застосування методи пошуку за ключовими словами, створення індексних словників, використання класифікаційних схем, тематичних рубрикаторів тощо. Це дає нам змогу стверджувати, що спеціалісти з інформаційних технологій аналогічно оцінили важливість методів визначення значущості документів для ранжування результатів пошуку. В інформаційно-пошукових системах бібліотек таку оцінку можна проводити на основі аналізу використання бібліотечних ресурсів.

Ранжування результатів інформаційного пошуку пропонується здійснювати на основі багатокритеріальної оцінки релевантності знайдених документів, що враховує наступні критерії:

– фактори старіння науково-технічної літератури (вік знайденого видання);

– значущість автора (індивідуального або колективного) на основі даних файлів авторитетних записів;

– статистичні дані про попит на літературу в бібліотеці.

Вибір критеріїв запропоновано з погляду на засади теорій бібліометрії та інформетрії.

Фактори старіння науково-технічної літератури

Старіння документів — це закономірний постійний процес зменшення з часом необхідності її використання для отримання вміщеної в ній інформації [9].

Стосовно до документів старіння розуміють не як фізичне старіння носія інформації, а як досить складний процес старіння інформації, що міститься в ньому. Зовні цей процес виявляється у втраті вченими і фахівцями інтересу до публікацій зі збільшенням часу, що пройшов з дня їхнього видання. Як показало обстеження 17 бібліотек, проведене одним з галузевих органів інформації, 62 % звертань приходиться на журнали, вік яких не перевищує 1,5 року; 31 % звертань — на журнали віком 1,5–5 років; 6 % — на журнали віком від 6 до 10 років; 7 % — на журнали більш ніж 10-літнього віку [10]. До публікацій, які вийшли порівняно давно, звертаються набагато рідше, що дає привід для ствердження про їхнє старіння.

Складність виявлення закономірностей старіння джерел інформації полягає в різниці характеристик зменшення їх використання в часі в різних предметних галузях і для різних часових періодів. Аналіз публікацій, присвячених теорії старіння інформації, дозволяє зробити висновок, що для апроксимації закономірностей старіння найчастіше використовуються від'ємні експоненціальні функції та від'ємні показникові функції [11].

Таким чином, фактори старіння літератури однозначно пов'язані з інтенсивністю її використання, і вік документа можна використовувати як критерій ранжування. Помилкою округлювання, викликаною методичними помилками техніки вимірювання можна знехтувати. Тоді вік обчислюється як різниця між поточним роком і роком публікації певного видання.

Значущість використання даного критерію для ранжування можливо підвищити, якщо для нормування числових значень критерію використовувати одну з функцій, що апроксимують закономірності старіння. Наприклад, для цього можна використати запропоновану ще в 1960 р. Р. Бартоном і Р. Кеблером функцію, що була призначена для апроксимації розподілу посилань, які були виявлені з пристатейної бібліографії у деяких галузях; часовий інтервал групування посилань — 10 років, кожен часовий інтервал позначається *x*; впорядкований за інтервалами і нормований потік посилань отриманий додаванням відносних величин мікропотоків послідовності часових інтервалів, позначених *y*. Запропонована ними залежність має вигляд функції:

$$y = 1 - \left(\frac{a}{e^{x}} + \frac{b}{e^{2x}} \right),$$

де *a* і *b* — емпіричні коефіцієнти, що відрізняються для різних галузей, причому $a + b = 1$.

Для апроксимації закономірностей старіння джерел інформації П. Коул у 1963 р. запропонував використовувати від'ємну експоненту наступного вигляду: [12]:

$$R_x = R_T e^{-\lambda x},$$

де R_x — кількість бібліографічних посилань старіше *x* років, що встановлюється шляхом підрахунку кумулятивного потоку від найстарішого віку до поточного моменту нульового року; R_T — максимальний кумулятивний потік посилань; λ — емпіричний коефіцієнт. Розрахунок емпіричного коефіцієнта П. Коул здійснював методом найменших квадратів, а щорічні пульсації мікропотоків нівелював збільшенням часових інтервалів до 5 років.

Крім наведених закономірностей можливо також використання закономірності Б. Брукса або інших з описаних в [11]. Всі ці моделі описують статистичні закономірності з достатньою точністю — відхилення теоретичних параметрів від емпіричних складають менше 0,09, що є досить припустимим.

Значущість автора на основі даних файлів авторитетних записів

У сучасних автоматизованих системах бібліотек, як відомо, використовуються спеціальні файли нормативних/авторитетних записів (Authority files). Ці записи створюються для деяких елементів бібліографічних записів: індивідуальних і колективних авторів, уніфікованих заголовків, заголовків серій, предметних рубрик, розділів класифікацій, дескрипторів тезаурусів. Таким чином, елементи нормативних записів виявляються при підготовці бібліографічних записів, а потім обробляються і доповнюються необхідними даними (варіантами, пов'язаними поняттями, іншими посиланнями і відсиланнями, текстами довідкового змісту, безліччю формальних ознак). Для представлення елементів таких записів використовуються формати типу UNIMARC і USMARC для Authority files. У практиці традицій-

ної каталогізації для таких елементів бібліографічних записів підготовляють додаткові картки або складають спеціальні картотеки, забезпечуючи уніфікацію опису видань і рукописів, повноту пошуку інформації.

Незважаючи на те, що авторитетні файли можна формувати для різних елементів заголовка, у світовій практиці, в першу чергу, виділяють такі з них, як: автор, колективний автор і предметна рубрика. Основним типом запису в авторитетному файлі є авторитетний запис, який представлений у машиночитній формі і містить інформацію про прийняті заголовки: авторський заголовок, уніфікований заголовок, предметна рубрика. Крім прийнятого заголовка в авторитетний запис включаються: варіантні форми заголовків, різні примітки й інша додаткова інформація, що характеризує заголовок або авторитетний запис у цілому. Сукупність авторитетних записів утворить авторитетний файл.

Для того, щоб використати інформацію з авторитетних файлів для ранжування результатів пошуку ІПС, потрібно, щоб дані бази авторитетних даних містили також певну якісну оцінку значущості автора. Питання введення такої оцінки потребує окремого дослідження, але деякі перспективи розвитку цього напрямку можна запропонувати вже зараз. По-перше, результатом автоматизованого семантичного аналізу авторитетних записів може стати ранжування (бальна оцінка) індивідуального або колективного автора. Так, для наукових бібліотек академічне видання буде мати найвищу оцінку, вузівське видання — дещо нижчу і так далі. По-друге, таку оцінку могли б проводити фахівці на етапі комплектування або створення авторитетного файлу.

Статистичні дані про попит на літературу в бібліотеці

Дані про використання літератури в бібліотеці можуть стати дуже цінним критерієм ранжування в пошуковій системі. Такі дані хоча і носять суб'єктивний характер, проте відображають цінність певного видання або автора. Інформація про використання літератури була корисною і бажаною для бібліотекарів і читачів з початку існування бібліотек. У період впровадження інформаційних технологій цінність її тільки підвищується, але питання одержання такої інформації все ще залишається актуальним. Бібліотеки, що ввели повний цикл автоматизованого обслуговування читачів отримують дані про використання літератури автоматично. Причому ця інформація є повною і багатоаспектною.

Нажаль, в Україні повністю автоматизоване обслуговування читачів є далекою перспективою. З огляду на це пропонується впровадження селективного моніторингу використання бібліотечних фондів для виявлення видань підвищеного попиту шляхом вибіркового аналізу відповідей книгосховища про незадоволені запити [13]. Це дасть змогу підготувати рекомендації для прийняття обґрунтованих управлінських рішень з оптимізації топології зберігання фондів та коригування політики комплектування літературою бібліотек.

Крім того, результати селективного моніторингу використання бібліотечних фондів можна використовувати для ранжування результатів пошуку в інформаційно-пошукових системах бібліотек. Недоліком використання запропонованої методики є певна похибка, викликана вибірковістю моніторингу, але цей недолік виправданий тим, що моніторинг усього інформаційного потоку потребує дуже

великих трудовитрат. Аналіз даних моніторингу дозволить сформувати репозиторій оперативних даних про використання бібліотечних фондів, який однозначно пов'язаний з електронним каталогом бібліотеки. На основі таких даних можна проранжирувати множину видань, отриману в результаті пошуку, згідно реального попиту на конкретне видання в певній бібліотеці.

Оскільки запропонована методика моніторингу використання бібліотечних ресурсів ґрунтується на аналізі відповідей книгосховища, то для коректного використання даних моніторингу в задачі ранжування результатів пошуку, як і в інших задачах, доцільне проведення аналізу отриманих даних для «згладжування» сплесків попиту, викликаних тимчасовими причинами, наприклад збільшенням попиту на учбову літературу в період сесії.

Числовими значеннями даного критерію є кількість вимог (відповідей книгосховища) на видання за певний період аналізування даних, наприклад за останній рік. Для нормалізації значень з метою використання критерію в задачі багатокритеріальної оцінки числові значення діляться на деякі нормуючі дільники, за які приймаються максимальні (мінімальні) значення критеріїв, що досягаються в області припустимих рішень

Використання цього критерію сумісно з наведеними вище дає змогу підвищити релевантність пошуку в інформаційно-пошукових системах бібліотек шляхом ранжування результатів пошуку.

Методика багатокритеріального ранжування результатів інформаційного пошуку

Задачу багатокритеріальної оптимізації можливо вирішити з використанням узагальненого (інтегрального) критерію.

Суть даного методу полягає в тому, що приватні критерії певним чином поєднуються в один інтегральний критерій, а потім знаходиться максимум або мінімум даного критерію.

Якщо об'єднання часткових критеріїв провадиться, виходячи з об'єктного взаємозв'язку часткових критеріїв і критерію узагальненого, то тоді оптимальне рішення буде коректне. Але таке об'єднання здійснити вкрай складно або неможливо, тому, як правило, узагальнений критерій є результатом чисто нормального об'єднання часткових критеріїв.

Лінійна згортка (адитивний критерій) використовується в моделях, заснованих на неявному постулаті: «низька оцінка за одним критерієм може бути компенсована високою оцінкою за іншим». У випадку ранжування результатів пошуку такий підхід прийнятний і тому використання лінійної згортки цілком обґрунтоване.

Цією моделлю користуються в задачах, у яких критерії мають ту саму одиницю виміру (наприклад, вартісну). Якщо критерії C_i не виражаються в тих самих одиницях виміру, то їх приводять до безрозмірного виду шляхом поділу значення кожного критерію на нормуючий коефіцієнт, що дорівнює максимальному значенню шкали для *i*-го критерію або за формулою:

$$Q = \frac{C_i - C_i^{\min}}{C_i^{\max} - C_i^{\min}},$$

де $C_i^{\min}$ і $C_i^{\max}$ — відповідно мінімальне і максимальне значення критеріїв якості на припустимій множині.

Потрібно зазначити, що в проблемі критеріального упорядкування альтернатив найтонше місце — це визначення вагових коефіцієнтів критеріїв. Теоретично, якісний аналіз важливості критеріїв може бути проведений різними шляхами. Один з найбільш доступних та змістовно обґрунтованих — застосування методів експертних оцінок.

Складність досліджуваного алгоритму полягає в тому, що важливість критеріїв має дуже суб'єктивний характер. Кожен користувач повинен мати можливість ранжирувати результати пошуку за будь-якою комбінацією критеріїв. Така можливість може бути надана за допомогою ітеративного інтерфейсу з використанням паралельно-послідовної стратегії пошуку. Тоді на певному кроці користувач зможе обрати ті критерії, за якими він хотів би відранжирувати знайдені документи. Крім того, він може відмовитись від ранжування, тоді результати пошуку будуть відсортовані за алфавітом.

Отже, у випадку використання такої технології, кількість критеріїв і їх важливість будуть визначатися динамічно.

Розглянемо запропоновану методику за алгоритмом.

Крок 1. Результати, отримані на першому етапі роботі ІПС, формують множину варіантів, які підлягають аналізу: $V = \{v_1, v_2,..., v_n\}$.

Крок 2. Сформуємо множину критеріїв, за якими оцінюються варіанти: $C = \{c_1, c_2,..., c_m\}$. Розробники ІПС повинні обрати один із двох можливих варіантів.

1 варіант. Критерії оцінки визначені заздалегідь і однакові для всіх користувачів у всіх сесіях пошуку.

2 варіант. Критерії оцінки визначаються динамічно, тобто користувач ІПС за допомогою ітеративного інтерфейсу на основі паралельно-послідовної стратегії пошуку обирає ті критерії, за якими він бажає проранжирувати отриману множину результатів пошуку.

Крок 3. Нормалізуємо числові значення критеріїв. Для цього числові значення часткових критеріїв діляться на деякі нормуючі дільники, за які приймаються максимальні (мінімальні) значення критеріїв, що досягаються в області припустимих рішень. Для тих, критеріїв, у яких оптимальне значення варіанту визначається мінімальним числовим значенням критерію (вік видання і значущість автора), нормалізоване значення визначається за формулою:

$$C_i^0 = \frac{C_i - C_i^{\min}}{C_i^{\max} - C_i^{\min}}, \qquad i = 1, ..., n,$$

де n — кількість варіантів з множини V.

Для тих, критеріїв, у яких оптимальне значення варіанту визначається максимальним числовим значенням критерію (дані про попит на видання в бібліотеці), нормалізоване значення визначається за тією ж формулою, але має протилежний знак:

$$C_i^0 = -\frac{C_i - C_i^{\min}}{C_i^{\max} - C_i^{\min}}.$$

Крок 4. Визначаємо вагові коефіцієнти (важливість) критеріїв β_j. На цьому кроці також можливі 2 варіанти.

1 варіант. Величини β_j для кожного критерію є постійними для певної комбінації критеріїв, причому для кожної комбінації $\sum_{j=1}^{m} \beta_j = 1$. Вагові коефіцієнти визначаються за допомогою методу експертних оцінок. Для оцінювання результатів вважаємо доцільним використання методу ранжування варіантів з оцінкою узгодженості думок експерта на основі коефіцієнта конкордації.

2 варіант. Важливість кожного критерію визначається безпосередньо користувачем для кожної сесії пошуку. Цей варіант є ідеальним для забезпечення інформаційної потреби конкретного користувача, але його реалізація є дуже важкою. Дане питання потребує подальшого дослідження.

Крок 5. Визначаємо значення інтегрального критерію для V_i, $i = 1,\ldots, n$:

$$C_i^{\text{int}} = \sum_{j=1}^{m} \beta_j C_{ij}^0,$$

де m — кількість поєднуваних часткових критеріїв; β_j — ваговий коефіцієнт j-го часткового критерію; C_{ij}^0 — нормоване значення j-го часткового критерію.

Крок 6. Упорядковуємо множину V за інтегральним критерієм C^{int}. Найкращим варіантом є той, для якого $C^{\text{int}} \to \min$. У випадку, коли значення C_{int} співпаде для різних варіантів з множини V, впорядкування проводиться за алфавітом.

Висновки

Важливим напрямком інтелектуалізації інформаційно-пошукових систем науково-технічних бібліотек є організація зворотного зв'язку в системах на основі використання статистичних даних для оцінки релевантності результатів пошуку. Необхідною умовою реалізації такого апарату є, по-перше, забезпечення механізму відстеження й аналізу результатів пошуку, і, по-друге, використання технології ранжування результатів пошуку за релевантністю знайдених документів. Використання запропонованої методики багатокритеріального ранжування результа-

тів інформаційного пошуку дозволить підвищити якість задоволення інформаційних потреб користувачів інформаційно-пошукових систем бібліотек.

1. *Роштейн А.П.* Интеллектуальные технологии идентификации: нечеткие множества, генетические алгоритмы, нейронные сети / Винницкий гос. технический ун-т. — Вінниця: УНІВЕРСУМ – Вінниця, 1999. — 302 с.

2. *Степанов М.Ф., Брагин Т.М.* Искусственные нейронные сети и их использование в интеллектуальных системах: Учебное пособие по дисциплине «Искусственные нейронные сети и их использование в интеллектуальных системах управления» для студ. направления 550200 / Саратовский гос. технический ун-т. — Саратов, 2000. — 126 с.

3. *Золкіна Е.А.* Моделі та методи оцінювання і розпізнавання двовимірних зображень в інтелектуальній діяльності людини: Дис… канд. техн. наук: 05.13.23 / Донецький держ. ін-т штучного інтелекту. — Донецьк, 2003. — 186 с.

4. *Андон Ф.И., Боровая Э.Н.* Методы представления и обработки некорректных данных и знаний в интеллектуальных информационных системах. — К., 1995. — 29 с.

5. *Боровая Э.Н.* Разработка методов представления и обработки некорректных данных и знаний в интеллектуальных информационных системах: Дис... канд. физ.-мат.наук: 05.13.11 / АН Украины; Ин-т программных систем. — К., 1994. — 143 с.

6. *Бень А.П.* Методы построения интеллектуальных адаптивных интерфейсов «человек – компьютеризированная система» на основе модели пользователя: Дис... канд. техн. наук: 05.13.06 / Херсонский гос. технический ун-т. — Херсон, 2000. — 194 с. — Бібліогр.: С. 155–174.

7. *Зайцева С.В.* Розробка прикладних систем з етапами модифікації та інтелектуалізації інтерфейсу користувача: Автореф. дис… канд. фіз.-мат. наук: 01.05.03 / НАН України; Ін-т кібернетики ім. В.М.Глушкова. — К., 2000. — 16 с.

8. *Харченко А.В.* Интеграция методов анализа больших наборов данных в интеллектуальном интерфейсе пользователя: Дис… канд. физ.-мат. наук: 01.05.03 / НАН Украины; Ин-т кибернетики им. В.М.Глушкова. — К., 1999. — 130 с. — Бібліогр.: С. 110–117.

9. *Мотылев В.М.* Старение научно-технической литературы. — Л., Наука, 1986. — 160 с.

10. *Чурсин Н.Н.* Популярная информатика. — К.: Техника, 1982.

11. *Горькова В.И.* Информетрия (Количественные методы в научно-технической информации) // Итоги науки и техники. Сер. Информатика. Т. 10. — М.: ВИНИТИ, 1988. — 328 с.

12. *Cole P.F.* Journal usage versus age of journal // J. Doc. — 1963. — Vol. 19, № 1. — P. 1–10.

13. *Яковлєва Ю.В.* Селективний моніторинг використання бібліотечних ресурсів // Реєстрація, зберігання і оброб. даних. — 2002. — Т. 4, № 1. — С. 89–96.

Надійшла до редакції 18.06.2004