УДК 022.4.001.57

В. В. Хаджинов[1], Ю. В. Яковлєва[2]
[1]Інститут проблем реєстрації інформації НАН України
вул. М.Шпака, 2, 03113 Київ, Україна
[2]Національна бібліотека України ім. В.І. Вернадського
пр. 40-річчя Жовтня, 3, 03039, Київ, Україна

Адаптивний пошук як напрям розвитку інформаційно-пошукових систем наукових бібліотек

Проаналізовано традиційні методи адаптації інформаційного пошуку, виділено методи, які доцільно використовувати в інформаційно-пошукових системах наукових бібліотек. Запропоновано напрями використання методу адаптації пошуку на основі модифікації запиту користувача.

Ключові слова: *інформаційно-пошукові системи, адаптивний пошук, автоматизація бібліотечних процесів, наукові бібліотеки.*

Збільшення обсягів інформаційних ресурсів, що надаються користувачам локальних і онлайнових інформаційно-пошукових систем бібліотек, потребує покращення таких показників пошуку, як повнота, точність, оперативність, зручність. Цій проблемі присвячено багато наукових праць, основна кількість яких припадає на кінець 70-х — початок 80-х років. Значний внесок у розробку теоретичних і прикладних питань підвищення ефективності інформаційного пошуку здійснили Г. Селтон, Дж. Солтон, Є.Ф. Скороходько, Е.Р. Сукіасян, В.В. Сидоренко, В.М. Дріянський, О.Г. Дубінський, Ю.В. Рогушина та ін. [1–3]. У зв'язку з розвитком Інтернет дослідження цих питань залишаються актуальними, зокрема, перспективною вважається розробка методів адаптивного пошуку.

У даній роботі розглянуто методи адаптації інформаційного пошуку й намічено ті методи, які доцільно використовувати в інформаційно-пошукових системах наукових бібліотек.

Основним критерієм ефективності пошуку інформаційно-пошукової системи (ІПС) бібліотеки є точність відповідності знайденого набору даних заданим критеріям пошуку. При цьому множини критеріїв, суттєві для певного користувача, можуть відрізнятися аж до зворотних. Отже, на нашу думку, інформаційно-пошукова система може вважатися ефективною, якщо вона включає засоби адаптивного пошуку, тобто здатна змінювати свої функціональні можливості або інтерфейс залежно від змінюваного в часі набору вимог користувачів.

Спираючись на роботи [4–11], можна виділити такі традиційні методи адаптації інформаційного пошуку до інформаційних потреб користувачів:

1) залучення людей-посередників, що уточнюють запит, використовуючи свої професійні знання;

2) розробка систем, у яких будуються моделі (профілі) користувача, що містять інформацію про його індивідуальні особливості, і використовувані для уточнення інформаційних користувацьких потреб;

3) автоматичні та напівавтоматичні операції із запитами (модифікація, розширення, зміна ваги термінів, ітеративне уточнення запиту тощо).

Перший спосіб є дуже ефективним, але вимагає залучення високооплачуваних фахівців, які зможуть обробляти одночасно тільки невелику кількість запитів.

У методах, що відносяться до другого способу, під моделюванням користувача в задачах інформаційного пошуку та фільтрації розуміють будь-який вид опису його інформаційних потреб для уточнення запитів.

Дослідження в сфері моделювання користувачів інформаційних систем у нашій країні ведуться ще з початку 70-х років минулого століття в роботах таких відомих учених як В.М. Глушков, А.М. Довгялло, Т.А. Гаврилова, В.І. Дракін та ін. Але розробка моделей користувачів для адаптації інформаційного пошуку не проводилася, хоча моделі користувачів (так звані «профілі користувачів») в ІПС використовувалися для виконання вибіркового поширення інформації [1].

За рубежем, як відзначено в [4, 5], перші роботи з моделювання користувачів комп'ютерних систем відносяться до кінця 70-х років. З того часу було розроблено численні прикладні системи, зібрано різні типи інформації про їхніх користувачів і визначено різні види адаптації до них. Однак, як відзначено в [6], більшість моделей користувачів створено в спеціалізованій формі, що не придатна для інших систем чи галузей. Ця спеціалізація створює труднощі в спільному використанні користувацьких моделей як загальних ресурсів для розробки систем інформаційного пошуку. Через це знання моделей користувачів, отриманих у специфічній області, не може бути узагальнене для більш широких цілей.

Нарешті методи третього способу адаптації пошуку передбачають модифікацію (зокрема — розширення) запитів — відомий і розповсюджений прийом в інформаційному пошуку. Модифікація запитів використовується для підвищення ефективності пошуку (тобто поліпшення таких показників як повнота або точність), а також для зручності користувача.

Методи модифікації запитів умовно можна поділити на 3 типи:

1) методи автоматичної модифікації запитів;

2) ітеративний інформаційний пошук, при якому користувач повторно формулює запит відповідно до результатів попереднього запиту, виданих пошуковою системою;

3) методи, засновані на аналізі колекції документів або тієї її частини, що видається у відповідь на первинний запит.

До першого типу відносяться методи лінгвістичної обробки запитів, а також модифікація запитів у формі природно-мовних питань [7]. Такого роду модифікації не зводяться до простого видалення ключових чи стоп-слів, і можуть бути досить витонченими. Наприклад, пошуковий сервер Yandex модифікує запити, що відповідають шаблонам «Що таке:?» і «Як зробити:?». ІПС, що «розуміють пи-

тання», дозволяють користувачеві висловлювати інформаційну потребу в більш природній формі.

Другий тип включає всі методи діалогового пошуку, в яких від користувача не вимагається оцінити релевантність знайдених результатів, проте надаються всі можливості модифікації запиту (уточнення запиту, вибір типу ранжирування знайдених документів тощо).

До останнього типу методів із запропонованої класифікації відноситься велика кількість перспективних методів пошуку, деякі з яких уже дуже поширені.

Один із традиційних методів розширення запитів — зворотний зв'язок по релевантності (*relevance feedback*) [8]. У рамках цього методу пошук розглядається як ітеративний процес, на кожному етапі якого відбувається уточнення інформаційної потреби користувача. Уточнення відбувається за рахунок того, що користувач явно вказує релевантні і нерелевантні документи в черговій видачі, що веде до модифікації запиту. Спрощеною реалізацією цього методу є функція «знайти схожі документи», представлена на багатьох пошукових системах. Як розвиток методу можна розглядати підхід, що використовує зворотний зв'язок по релевантності на етапі навчання системи переформулювання запитів. Після навчання система розширює запити без участі користувача.

Інший підхід — розширення запитів на основі статистики спільного входження слів у всій колекції або в окремій видачі [8]. Так, наприкінці 90-х років минулого століття пошукова система AltaVista (www.altavista.com) надавала сервіс *AltaVista Refine*, що дозволяв уточнювати запит за допомогою словника спільного входження слів [9].

Також існують методи на основі спеціальних словників — тезаурусів. Традиційно в інформаційному пошуку для розширення запитів використовуються семантичні словники — тезауруси [2]. Тезауруси можуть бути побудовані автоматично на основі аналізу спільного входження слів, а також вручну. На початковому етапі розвитку інформаційного пошуку тезауруси служили для стандартизації словника ІПС та економії пам'яті. Згодом основною функцією тезаурусів стало підвищення повноти пошуку за рахунок об'єднання синонімічно й семантично близьких термінів по OR. Методи розширення запитів за допомогою тезаурусів широко обговорювалися в літературі, повідомлялися суперечливі результати. Однак побудовані вручну тезауруси, звичайно, дають гарні результати в різних задачах [10, 11].

Специфіка ІПС наукових бібліотек дозволяє використовувати багато з наведених методів адаптації пошуку. Ці методи вже активно досліджуються й впроваджуються в багатьох бібліотеках, хоча сьогодні залишається ще багато невирішених задач.

На нашу думку доцільним є дослідження й використання методу адаптації пошуку на основі модифікації запиту, зокрема, такого, який передбачає організацію зворотного зв'язку в системах на основі використання статистичних даних для оцінки релевантності результатів пошуку. Зворотний зв'язок по релевантності пропонується реалізувати за допомогою двох засобів.

Першим є забезпечення механізму відстеження й аналізу результатів пошуку, тобто статистичний аналіз запитів користувачів. Зворотним зв'язком від користувача при цьому є непряма оцінка релевантності знайдених документів, що спира-

ється на відбір користувачем документів для автоматичного замовлення або для збереження в підмножину чи файл (у випадку часткової автоматизації бібліотечних процесів).

Другим, але більш важливим засобом є використання запропонованої нами в [12] технології ранжирування результатів пошуку, що базується на багатокритеріальній оцінці видань за інформативністю.

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

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

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

— статистичні дані щодо обігу літератури в бібліотеці.

Вибір критеріїв запропонований із погляду на засади теорій бібліометрії та інформетрії. На нашу думку, наведені критерії якнайкраще визначають інформативність документів [13]. Перший критерій — старіння літератури, як закономірний постійний процес зменшення із часом необхідності її використання для отримання вміщеної в ній інформації, однозначно пов'язаний з інтенсивністю використання документів.

Використання другого критерію (даних файлів авторитетних записів) можливе за умов розробки методів оцінки значущості авторів наукових публікацій. Перспективним розвитком цього напряму може стати автоматичний семантичний аналіз авторитетних записів, результатом якого буде ранжирування (бальна оцінка) індивідуального або колективного автора. Так, академічне видання матиме найвищу оцінку, вузівське — дещо нижчу, і так далі. Питання введення такої оцінки потребує окремого дослідження за участю спеціалістів з інформаційних технологій та бібліотекознавців-практиків.

Не викликає сумнівів цінність третього критерію, а саме даних системи моніторингу використання бібліотечних фондів у науковій бібліотеці. У роботі [14] нами запропоновано методику селективного моніторингу для виявлення видань підвищеного попиту шляхом вибіркового аналізу відповідей книгосховища про незадоволені запити. Числовими значеннями даного критерію є кількість вимог (відповідей книгосховища) на видання за визначений період аналізування даних.

Саме цей критерій дозволить реалізувати зворотний зв'язок в ІПС, тобто оцінити релевантність знайденого документа на основі статистики його використання в бібліотеці.

Задачу багатокритеріальної оптимізації можливо вирішити, застосувавши інтегральний критерій. Лінійна згортка (або інтегральний критерій) використовується в моделях, заснованих на неявному постулаті: «Низька оцінка за одним критерієм може бути компенсована високою оцінкою за іншим». У випадку ранжирування результатів пошуку такий підхід прийнятний і тому використання лінійної згортки цілком обґрунтоване.

Значення інтегрального критерію визначається для кожного документа v_i з множини варіантів $V = \{v_1, v_2, ..., v_n\}$, які підлягають аналізу й отримані на першому етапі роботи ІПС:

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

де m — кількість поєднуваних часткових критеріїв; $i = 1,\ldots, n$ — номер варіанта (n — кількість варіантів із множини V); β_j — ваговий коефіцієнт j-го часткового критерію; C_{ij}^0 — нормоване значення j-го часткового критерію.

Для нормалізації числові значення часткових критеріїв діляться на деякі нормуючі дільники, за які приймаються максимальні (мінімальні) значення критеріїв, що досягаються в області припустимих рішень [13]. Для критеріїв, у яких оптимальне значення варіанта визначається мінімальним числовим значенням критерію (у нашому випадку це вік видання й значущість автора), нормалізоване значення визначається за формулою:

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

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

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

Потрібно зазначити, що в проблемі критеріального впорядкування альтернатив найтонше місце — це визначення вагових коефіцієнтів критеріїв. Один із найбільш доступних та змістовно обґрунтованих способів якісного аналізу важливості критеріїв — застосування методів експертних оцінок.

Складність досліджуваного алгоритму полягає в тому, що важливість критеріїв має дуже суб'єктивний характер. Тому доцільно (хоч і важче в реалізації та користуванні) надати кожному користувачеві можливість ранжирувати результати пошуку за будь-якою комбінацією критеріїв. Така можливість може бути надана за допомогою адаптивного інтерфейсу, тоді на певному кроці користувач зможе обрати критерії, за якими він хотів би відранжирувати знайдені документи, і визначити їхню пріоритетність, а саме присвоїти кожному критерію ранг (номер), що підвищується зі спадом значущості критерію. Крім того, він може відмовитись від ранжирування, тоді результати пошуку будуть відсортовані за алфавітом.

Визначення користувачем ступеню пріоритетності, тобто ваги критерію, дуже ускладнює роботу з ІПС, тому ми вважаємо просте впорядкування критеріїв цілком достатнім. Тоді, враховуючи, що $\sum_{j=1}^{n} \beta_j = 1$, для кожного випадку вагові коефіцієнти пропонуємо визначати у такий спосіб:

— якщо обрано 2 критерії, то коефіцієнт першого з них дорівнюватиме 2/3, а

другого — 1/3;

— якщо обрано 3 критерії, то коефіцієнти дорівнюватимуть 1/2, 1/3 і 1/6 відповідно.

У системі зі зворотним зв'язком по релевантності критерій, що враховує попит на видання, буде визначатися як:

$$C_i^0 = \beta_s C_{is}^0 + \beta_p C_{ip}^0, \tag{2}$$

де $i = 1,\ldots, n$ — номер варіанта з множини V; C_{is}^0 — нормоване значення критерію, що визначає попит на основі оцінки користувачем релевантності документа (зворотний зв'язок в ІПС); C_{ip}^0 — нормоване значення критерію, що визначає попит на основі даних системи моніторингу бібліотечних фондів; β_s і β_p — відповідні вагові коефіцієнти часткових критеріїв.

Головним призначенням критерію (2) є його використання в системі моніторингу, при цьому значення C_{is}^0 надходить від ІПС як додаткова інформація про попит на певні видання. Вагові коефіцієнти в цьому випадку доцільно визначити за допомогою експертної оцінки.

Отже, підставивши (2) в (1), отримаємо вираз для визначення інтегрального критерію у випадку використаннях трьох критеріїв оцінки інформативності документа:

$$C_i^{\text{int}} = \beta_1 C_{i1}^0 + \beta_2 C_{i2}^0 + \beta_3 \left(\beta_s C_{is}^0 + \beta_p C_{ip}^0\right). \tag{3}$$

Очевидно, що під третім критерієм у (3) мається на увазі попит на документ.

Отже, наведені нами засоби адаптивного пошуку дозволять реалізувати двосторонній зв'язок між ІПС і системою моніторингу використання бібліотечних ресурсів (СМ), зокрема, ІПС використовуватиме дані системи моніторингу бібліотечних фондів для ранжирування результатів пошуку, а СМ буде враховувати дані системи відстеження й аналізу запитів користувачів як додаткову інформацію про попит на певні видання (див. рисунок). Такий зворотний зв'язок є досить логічним, оскільки на етапі часткової автоматизації бібліотек запити до локальних і онлайнових ІПС бібліотек виявляються єдиними електронними документами на шляху вимог, і їхній аналіз підвищить ефективність моніторингу бібліотечних фондів.

Такі зв'язки між ІПС і СМ дозволять, по-перше, покращити якість пошуку в ІПС бібліотек шляхом впровадження засобів адаптивного пошуку, і, по-друге, підвищити ефективність системи моніторингу використання фондів, спрямованої на надання рекомендацій для прийняття обґрунтованих управлінських рішень щодо оптимізації топології зберігання фондів та коригування політики комплектування літературою бібліотек.

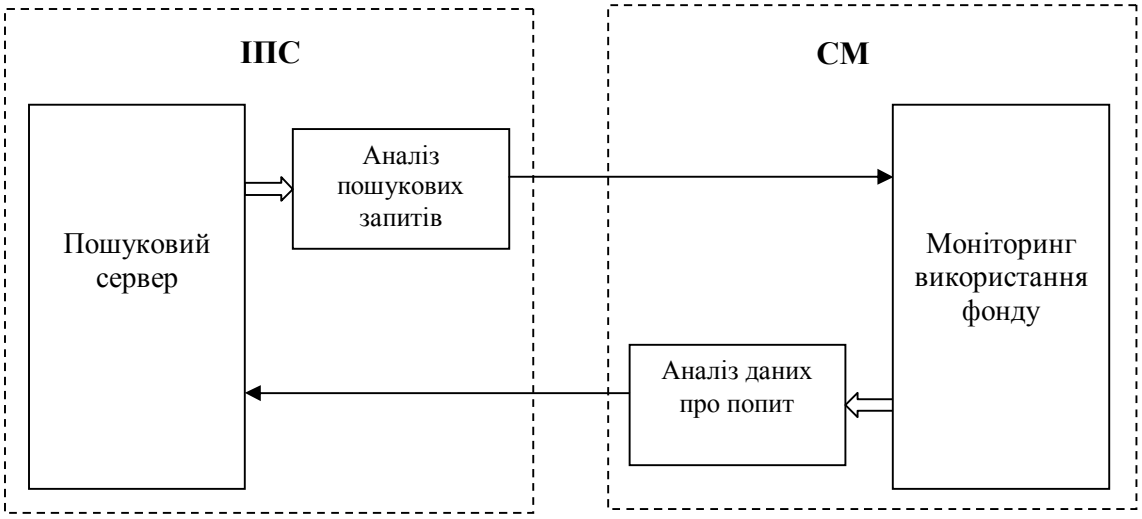

Схема взаємозв'язку між ІПС та СМ.

Висновки

Впровадження засобів адаптивного пошуку в інформаційно-пошукових системах бібліотек є досить перспективним. Зокрема, доцільним є дослідження методів модифікації запитів, що передбачають реалізацію зворотного зв'язку в ІПС на основі оцінки релевантності документів. Саме ІПС бібліотек мають унікальну можливість крім застосування методів, що використовуються пошуковими серверами Інтернет, також оцінювати релевантність документів на основі статистичного аналізу попиту на певні видання в бібліотеці. Запропонована методика багатокритеріального ранжирування результатів інформаційного пошуку дозволить підвищити якість задоволення інформаційних потреб користувачів інформаційно-пошукових систем наукових бібліотек.

1. *Ночевнов Д.П.* Методи та засоби адаптивного інформаційного пошуку на основі моделі користувача: Дис. ... канд. техн. наук: 05.13.06 / Черкас. держ. технол. ун-т. — Черкаси, 2005. — 178 с.

2. *Солтон Дж.* Динамические библиографические системы: Пер. с англ. / Под ред. В.Р.Хисамутдинова. — М.: Мир, 1979. — 557 с.

3. *Сукиасян Э.Р.* Homo Quaerens (Человек ищущий). К проблеме развития познавательных способностей читателя в процессе информационного поиска / Э.Р.Сукиасян // НТБ. — М., 2002.— № 4. — С. 32–33.

4. *Belkin N.J.* User Modelling in Information Retrieval // Tutorial presented in Sixth International Conference on User Modelling. — Chia Laguna (Sardinia). — 1997, June.

5. *Jameson A.* User Adaptive Systems: An Integra-tive Overview // Tutorial originally presented at UM99 (June, 1999) and IJCA199 (August, 1999). Department of Computer Science. — Saarland University. — Germany.

6. *Ночевнов Д.П.* Системный анализ методов адаптации информационного поиска в информационно-поисковых системах // Вісн. Черкас. держ. технол. ун-ту. — Черкаси, 2003.

7. *Альшанский Г.В., Браславский П.И., Титов П.В.* Формирование информационных запросов к машинам поиска интернета на основе тезауруса [Электрон. ресурс] // Материалы VIII Междунар. конф. по электрон. публикациям «EL-Pub2003», 8–10 октября 2003 г., — Новосибирск — Способ доступа: URL: http://www.ict.nsc.ru/ws/elpub2003/5964/. — Загл. с экрана.

8. *Baeza-Yates R., Ribeiro-Neto B*. Modern Information Retrieval. — New York et al.: ACM Press, Addison-Wesley, 1999. — 513 p.

9. *Schwarz C.* Web Search Engines // Journal of the American Society for Information Science.– 1998. — 49 (11). — P. 973–982.

10. *Браславский П.И.* Автоматические операции с запросами к машинам поиска интернета на основе тезауруса: подходы и оценки [Электрон. ресурс] // Труды Международного семинара Диалог-2004 по компьютерной лингвистике и ее приложениям. — М., 2004. — Способ доступа: URL: http://www.dialog-21.ru/Archive/2004/Braslavskij.htm. — Загл. с экрана.

11. *Bodner R., Song F.* Knowledge-Based Approaches to Query Expansion in Information Retrieval. In McCalla, G. (Ed.). — Advances in Artificial Intelligence. — New York: Springer, 1996. — P. 146–158.

12. *Яковлєва Ю.В.* Методика ранжирування результатів пошуку в інформаційно-пошукових системах бібліотек // Реєстрація, зберігання і оброб. даних. — 2004. — Т. 6, № 3. — С. 66–73.

13. *Яковлєва Ю.В.* Оцінка інформативності документів у пошукових системах наукових бібліотек // НТІ. — 2004. — № 4. — С. 52–54.

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

Надійшла до редакції 13.04.2006