УДК 681.3

Д. В. Ланде
Інститут проблем реєстрації інформації НАН України
вул. М. Шпака, 2, 03113 Київ, Україна

Шляхи доступу до документальних інформаційних ресурсів у веб-просторі

Розглянуто технології та шляхи доступу користувачів до документальних ресурсів у веб-просторі. Наведено рекомендації щодо створення спеціалізованої мережевої метапошукової системи.

Ключові слова: *інформаційно-пошукова система, метапошукова система, документ, інформаційне джерело, веб-простір, веб-ресурси.*

На сьогоднішній день підтримка інформаційно-аналітичної діяльності шляхом застосування методів і засобів моніторингу, адаптивного агрегування та узагальнення потоків інформації з глобальних комп'ютерних мереж, насамперед мережі Інтернет і відповідних оверлейних мереж, є актуальною науково-практичною проблемою [1, 2].

Розвиток інформаційно-комунікаційних технологій призвів до різкого зростання обсягів інформації, яка напрацьовується, зберігається, обробляється, нагромаджується та поширюється в інформаційному просторі. У той же час своєчасне одержання багатоаспектної та об'єктивної інформації з комп'ютерних мереж для подальшого її використання у різноманітній аналітичній діяльності потребує застосування сучасних технологічних рішень.

Починаючи роботу в Інтернеті, користувачі спочатку звертаються до вибраних джерел, постійно відслідковують (здійснюють моніторинг) зміни, динаміку появи нових матеріалів (***алгоритм 1***). Наступним етапом використання ресурсів веб-простору зазвичай є застосування мережевих інформаційно-пошукових систем (ІПС), кожна з яких має свої особливості, але «монополістом» серед яких є, звичайно, Google (***алгоритм 2***). Після цього користувач звертається до спеціалізованих мережевих метапошукових систем (***алгоритм 3***), які агрегують можливості звичайних мережевих ІПС, деякі з яких адаптовано під інформаційні потреби своїх користувачів.

Якщо ранжирувати кількість джерел, які можна отримати при застосуванні трьох наведених вище підходів, то, ймовірно, можна у черговий раз отримати підтвердження загальнонаукової закономірності Бредфорда, яка, у свою чергу витікає

із закону Ціпфа [3, 4]. Закономірність Бредфорда у початковому вигляді відносилася до традиційних «паперових» періодичних видань. Досліджуючи різні типи джерел інформації, Бредфорд розподілив їх за трьома множинами, рівними за кількістю релевантних документів: R_1, R_2, R_3, де R_1 — це найбільш рейтингові джерела, які безпосередньо відносяться до певної тематики; R_2 — множина джерел, що кореспондуються з комп'ютерною тематикою; R_3 — джерела, які лише частково торкаються даної теми. При цьому кількість корисної інформації в усіх трьох множинах є сталою.

Якщо прийняти позначення, що $|A|$ — це кількість елементів множини A, то пропорція Бредфорда записується у такий спосіб:

$$|R_1| : |R_2| : |R_3| = C.$$

Для множин документальних джерел, які отримуються за наведеними вище алгоритмами, відповідно, справедливо:

$$|S_1| : |S_3| = |S_3| : |S_2| = C,$$

де S_1 — множина джерел, що отримані за *алгоритмом 1* (вибрані джерела з веб-простору); S_2 — множина джерел, які кореспондуються з *алгоритмом 2* (пошук у глобальних мережевих ІПС); S_3 — джерела, що відповідають *алгоритму 3* (застосування спеціалізованих метапошукових систем); C — деяка константа, що відповідає інформаційним потребам користувачів.

Моніторинг вибраних мережевих джерел

Нині у веб-просторі знаходиться велика кількість інформаційних ресурсів, що представлені в документальних форматах (DOC, RTF, PS, SGML, HTML, PDF тощо). Найбільш вживаним із цих форматів є PDF, популярність якого викликана тим, що він є компактним і зручним для представлення та зберігання інформації, що наведена у вихідному стані в різному вигляді: простого тексту, векторних і растрових зображень, сторінок веб-сайтів, форм і мультимедійних файлів [4]. Саме у цьому форматі представлена документальна інформація на сучасних наукових і аналітичних веб-сайтах.

На сьогодні найпопулярнішим веб-ресурсом, на якому вченими всього світу публікуються препринти, є сервер Корнельського університету (США) ArXiv (http://www.arxiv.org — рис. 1). На даний час на цьому веб-ресурсі опубліковано понад 700 тис. електронних препринтів. Багато із статей з природничих наук (і вже не тільки) перед офіційною публікацією проходять «препринтову стадію» на цьому сервері [6]. Саме з моніторингу оновлень цього сайту мільйони вчених починають свій робочий день. До речі, інформаційні оновлення ArXiv вільно розповсюджуються електронною поштою у вигляді вихідних даних і переліків анонсів.

Безумовно, найбільшим джерелом бібліографічної інформації у світі є веб-сервер Бібліотеки Конгресу США (http://www.loc.gov/), більшість ресурсів якої, нажаль, є недоступними для вільного доступу в повному обсязі. Вченим з України та всього світу вільний доступ до української наукової періодики та монографій надала Національна бібліотека України ім. В.І. Вернадського (http://nbuv.gov.ua).

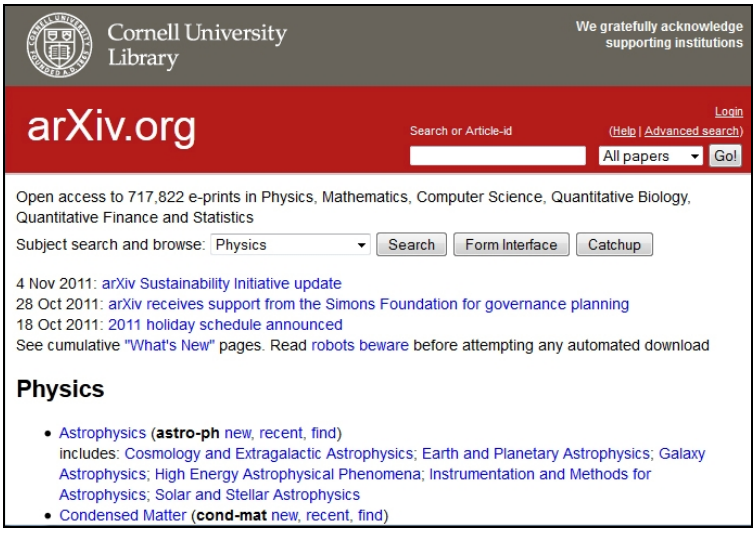

Рис. 1. Сайт ArXiv (http://www.ArXiv.org)

Користувачам-науковцям можна рекомендувати ще декілька веб-ресурсів, серед яких виділяються Єдине вікно, доступ до освітніх ресурсів (http://window.edu.ru), Наукова електронна бібліотека (http://elibrary.ru), електронний архів Російської академії природознавства (http://econf.rae.ru), інші системи відкритого доступу до наукової інформації (наприклад, http://mendeley.com, http://highwire.stanford.edu, http://sworld.com.ua).

Використання мережевих інформаційно-пошукових систем

Практично всі популярні мережеві інформаційно-пошукові системи в режимі розширеного доступу пропонують здійснити пошук документів у форматі PDF.

Наприклад, пошукові системи Google (http://google.com), Bing (http://bing.com) і Yahoo! (http://yahoo.com) надають можливість пошуку PDF-файлів шляхом додання до запиту користувача виразу: «filetype:pdf». При цьому відповіді двох останніх систем на пошукові запити повністю співпадають, що підтверджує підозру щодо використання ними спільних баз даних і пошукових машин.

У системі Яндекс для пошуку PDF-файлів необхідно безпосередньо звернутися до форми розширеного пошуку (http://yandex.ua/search/advanced — рис. 2).

Разом з тим, при пошуку необхідної документації у форматі PDF за допомогою традиційних мережевих інформаційно-пошукових систем користувач постійно стикається з проблемами, що пов’язані з поганою доступністю цільової інформації (умовами платного доступу, відсутністю необхідних файлів за вказаними адресами або невірними гіперпосиланнями). Хоча більшість пошукових систем, таких як Google, Яндекс, Bing, Yahoo виводять у список результатів інформацію щодо знайдених PDF-файлів, разом з тим вони часто дають посилання на неіснуючі PDF-файли або посилання на веб-сайти, де PDF-файли знаходяться у закритому доступі.

Ще одним ресурсом для пошуку документальної інформації компанії Google є Google Scholar (українська версія — Google Академія) — вільно доступна ІПС,

яка індексує повний текст наукових документів (рис. 3). Індекс Google Scholar охоплює більшість рецензуємих онлайн-журналів найбільших наукових видавництв. За своїми функціями Google Академія подібна до таких вільно доступних систем як Scirus от Elsevier, CiteSeerX и getCITED.

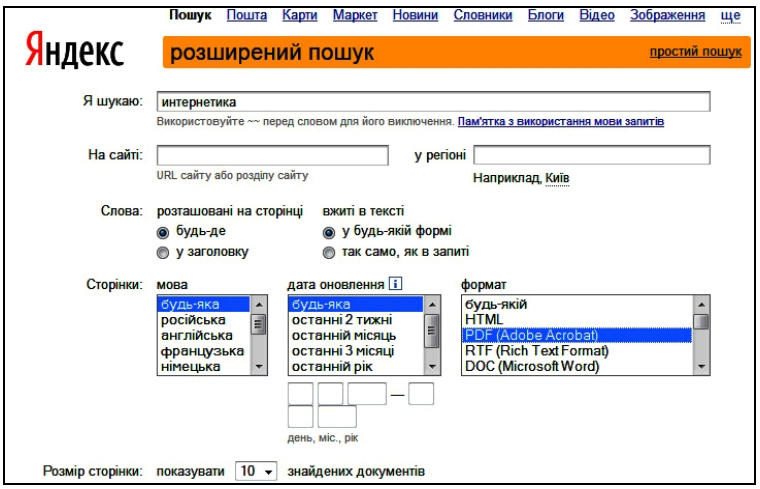

Рис. 2. Форма розширеного пошуку в системі Яндекс (http://yandex.ua/search/advanced)

Рис. 3. Google Scholar (http://scholar.google.com)

Google Академія ранжирує результати за допомогою комбінованого алгоритму, який надає великі переваги кількості цитат і слів, що входять до заголовку документа. Це веде до того, що перші результати пошуку містять найбільш цитовані документи.

Використання спеціальних метапошукових систем

Традиційна мережева інформаційно-пошукова система в процесі роботи переглядає певний набір серверів і відбирає документи відповідно до заданих кри-

теріїв. Звичайно, пошук за допомогою різних систем за одними й тими ж запитами дає різні результати. Це привело до ідеї створення так званих метапошукових систем, які звертаються за допомогою відразу до декількох пошукових систем. Результати пошуку з усіх систем об'єднуються і надаються користувачеві у відповідній формі. Природно, пошук за допомогою метапошукових систем займає більше часу порівняно зі звичайними ІПС.

Деякі з метапошукових систем, не маючи власних баз даних і власних розвинених пошукових механізмів, компенсують це можливостями аналізу, групування та візуалізації результатів пошуку. Наприклад, система Nigma.ru використовує технологію відображення первинних результатів пошуку у вигляді «хмари» тегів. Кожен тег можна «розкривати», внаслідок чого отримати відповідні підлеглі теги. Результати пошуку, відповідні вибраним тегам, відображуються у вигляді переліку.

Інша російська пошукова система Quintura (http://quintura.ru) має інтерфейс, який забезпечує отримання автоматичних підказок за введеним запитом, допомагає динамічно управляти процесом пошуку. За запитом з одного слова Quintura пред'являє можливі фрази та словосполучення, якими за необхідності розширюється первинний пошуковий запит.

Нині жодна з універсальних мережевих пошукових систем на достатньому рівні не допомагає при пошуку документальної інформації. Найкращими засобами на цей час виступають вузько спеціалізовані метапошукові системи, що орієнтовані саме на пошук документальної інформації, найчастіше представленої у форматі PDF.

Разом з тим, наведені вище недоліки традиційних інформаційно-пошукових систем (платний доступ до окремих документів, хибні гіперпосилання тощо) мають місце й для спеціалізованих метапошукових систем, що орієнтовані на пошук документів у форматі PDF, таких як OSUN (http://www.osun.org), PDFGod (http://www.pdfgod.com), PDF Search Engine (http://pdf-search-engine.com), PDF DataBase (http://pdfdatabase.com) тощо.

У наведених пошукових системах немає можливості сортування і фільтрації результатів пошуку, або просто пошуку в базі даних уже збережених PDF-файлів. Усі названі системи більшою частиною спрямовані на англомовних користувачів і використовують для отримання інформації в основному систему Google, що обмежує результати пошуку. Крім того, лише одна зі спеціалізованих пошукових систем PDF Search Engine може видавати PDF-файли у HTML-вигляді (для попереднього ознайомлення зі змістом документів).

Розроблена під керівництвом автора метапошукова система PDFSS (PDF Science Search) (http://weblib.in.ua — рис. 4.) з самого початку була створена як система пошуку науково-технічної документації і використовувалася користувачами, які шукали саме такі документи. Основна ідея цієї системи полягає у тому, щоб знаходити у веб-просторі PDF-файли без супроводжуючого їх інформаційного шуму або реклами.

Особливістю PDFSS є те, що вона повністю спрямована на пошук доступних користувачеві PDF-файлів, з можливістю фільтрації платних ресурсів, текстових описів, будь-якої інформації, окрім самих документальних файлів. У адаптивному кеші PDFSS присутні переважно науково-технічні документи (нині їхня кількість перевищує 500 000) з більш ніж 50 тис. джерел. Лідирують серед джерел для

PDFSS сайт nbuv.gov.ua (Національна бібліотека України ім. В.І. Вернадського), ioffe.ru (Фізико-технічний інститут ім. О.Ф. Іоффе), window.edu.ru (Єдине вікно, доступ до освітніх ресурсів) та ін. Більшість із джерел — це сайти університетів, інститутів, а також наукових журналів і електронних бібліотек.

Рис. 4. Інтерфейс метапошукової системи PDFSS

Метапошукова система PDFSS складається з трьох основних модулів:

— модуля метапошуку;

— модуля кешування інформації (інформаційний проксі-сервер);

— внутрішньої пошукової системи, яка працює як з інформаційним проксі-сервером, так і репозитарієм PDFSS, що створюється шляхом архівування даних з інформаційного проксі-сервера.

Основним критерієм ранжирування інформації в системі PDFSS є рейтинг пошукових систем. Так, наприклад, у пошукової системи Google рейтинг вищий, ніж у системи Bing (у Google більше охоплення ресурсів, більш релевантні результати). У PDFSS відбувається фільтрація неінформативних сайтів або сайтів з недоступними першоджерелами (так званий «стоп-перелік»).

Якщо посилання на PDF-документ було отримано з різних пошукових систем, то обирається те з них, яке містить найбільш повний опис. Результати надаються користувачеві у вигляді переліків результатів різних пошукових систем, які слідують один за одним.

У системі PDFSS використовується модуль кешування, основне завдання якого — збір посилань на PDF-документи, що отримані в процесі роботи користувачів, щоб надалі зберегти в інформаційному сховищі файли (адаптивне агрегування інформації), а також пов’язану з ними інформацію, таку як доступність файлів, розмір файлів. Внутрішня інформаційно-пошукова система дозволяє користувачеві шукати в кеші системи PDFSS документи, які динамічно накопичуються.

Порівняння результатів експлуатації системи PDFSS з іншими подібними системами дозволяє зробити висновок про її кращу орієнтацію на українську і російську мови.

Підходи до створення спеціалізованої метапошукової системи

Спеціалізована метапошукова система, модель якої пропонується, має базуватися на ідеології, що реалізована в системі PDFSS, але доповнена категоріальністю користувачів і застосуванням інтерактивних графічних інтерфейсів.

Загальна схема роботи нової метапошукової системи передбачає ряд етапів:

— після того, як користувач задасть запит, мають створитися запити для кожної з пошукових систем, які враховують унікальні можливості їхнього синтаксису;

— модифіковані запити пересилаються пошуковим системам, які повертають результати пошуку;

— метапошукова система розбирає отримані результати на окремі документи і перевіряє їхню доступність (наприклад, якщо в шляху до файлу є доменне ім'я, присутнє в «стоп-переліку», то файл відкидається і не використовується у подальшій обробці);

— здійснюється пошук у внутрішній базі цих файлів — інформаційному кеші, що містить знайдені раніше документи. (*Якщо такі файли вже були знайдені, то виведення документа доповнюється інформацією про можливу доступність цього файлу за знайденим посиланням. Якщо цей файл відсутній за вказаною адресою у веб-просторі, то виводиться повідомлення, що цей файл може бути відсутнім. Якщо ж інформація про файл присутня в інформаційному кеші, і він імовірно існує, то виведення доповнюється такою інформацією як розмір файлу, також пропонується HTML-версія цього файлу*);

— документи перетворяться для виведення результатів і виводяться користувачеві через стандартний веб-інтерфейс.

Передбачається три основні категорії користувачів:

— анонімний користувач;

— зареєстрований користувач;

— адміністратор.

Анонімний користувач повинен мати права мультипошуку документів у веб-просторі без доступу до інформаційного кешу системи.

Зареєстрований користувач має права анонімного користувача, крім того, він має права пошуку як у загальному інформаційному кеші, так і в масиві вибраних і збережених ним у кеші документів. Результати його пошуку враховуються системою, всі видані йому документи зберігаються у інформаційному кеші (але через визначений адміністратором час вилучаються), а вибрані документи зберігаються у кеші без часових обмежень. Зареєстрований користувач може розміщати й власні документи в інформаційний кеш системи (у свою область), крім того, він може зберігати та редагувати свої запити для їхнього подальшого використання.

Зареєстрований користувач повинен мати власну веб-сторінку на веб-сайті системи (персональний кабінет), де він може вводити або модифікувати дані про себе (логін, пароль, електронна адреса, ПІБ), рубрики інтересів, збережені запити.

Адміністратор має всі права зареєстрованого користувача, крім того він має право:

— поповнювати «стоп-перелік» сайтів, які вилучаються з результатів пошуку;

— вилучати будь-які документи з інформаційного кешу системи;

— вилучати користувачів, які, наприклад, порушують угоду щодо використання можливостей системи метапошуку.

Основу інтерфейсу користувачів з адаптивним документальним сховищем має скласти система смислової навігації, яка має бути побудована за допомогою сучасних програмних засобів візуалізації.

Висновки

Проаналізовано технологічні засади організації метапошукових систем у масивах документальної інформації у веб-просторі. Вибрані технологічні засади мають бути застосовані при побудові моделі метапошуку, яка орієнтована на роботу з документальною інформацією (статтями, тезами доповідей, дисертаціями, науковими звітами і т.п.) при проведенні інформаційно-аналітичної діяльності науковців, державних службовців, бізнес-аналітиків тощо.

Очікувані результати дозволять поєднати у єдиному технологічному ланцюжку інформаційний пошук зі змістовним аналізом даних, що підвищить якість опрацювання поточної інформації і відповідно ефективність інформаційної підтримки аналітичної діяльності.

1. *Додонов А.Г.* Современные поисковые технологии - проблемы и некоторые пути их решения / А.Г. Додонов, Д.В. Ланде, В.Г. Путятин // Реєстрація, зберігання і оброб. даних. — 2010. — Т. 12, № 3. — С. 36–55.

2. *Додонов А.Г.* Методы и средства мониторинга, адаптивного агрегирования и обобщения информационнных потоков / А.Г. Додонов, Д.В. Ландэ // Информационные технологии и безопасность. Проблемы научного и правового обеспечения кибербезопасности в современном мире. Материалы международной научной конференции ИТБ-2011. — К.: ИПРИ НАН Украины, 2011. — С. 6–9.

3. *Bradford S.C.* Sources of Information on Specific Subjects, Engineering // An Illustrated Weekly Journal (London). — 1934 (26 January). — **137**. — P. 85–86.

4. *Ландэ Д.В.* Поиск знаний в Internet. Профессиональная работа. — М.: «Вильямс», 2005. — 272 с.

5. *Document* Management. Portable Document Format. Part 1: PDF 1.7 // Adobe Systems Inc. — 2008. — 756 p. — http://www.adobe.com/devnet/acrobat/PDFs/PDF32000_2008.PDF

6. *Ландэ Д.В.* Отражение целевой тематики в публикациях и электронных препринтах ArXiv / Д.В. Ландэ, А.А. Снарский // Матеріали 16-ї Міжнар. наук.-техн. конф. «Електромагнітні та акустичні методи неруйнівного контролю матеріалів та виробів», 21–26 лютого 2011 р. — Славське Львівської області. — С. 74–77.

Надійшла до редакції 28.11.2011