УДК 025.4.036

Валерія Струнгар,
бібліотекар І категорії інформаційно-аналітичного відділу
Фонду Президентів України
Національної бібліотеки України імені В. І. Вернадського

Інформаційно-пошукова система бібліотеки як інструмент прискорення опрацювання даних

У пропонованій статті досліджуються різні питання розроблення і створення інформаційно-пошукових систем бібліотеки, які здатні в автоматизованому режимі отримувати дані із електронних документів з метою їхнього впровадження у науково-інформаційний процес.

Ключові слова: *інформаційний простір, комунікації, інформаційно-пошукова система, WWW-сервіс, технології, сайти, інтернет-ресурси.*

Постановка проблеми. Проблема пошуку і доступу до інформації є однією з серйозних проблем, з якою зіткнулося сучасне "інформаційне суспільство". Один із перших вчених, що найчіткіше усвідомив її, був бельгійський соціолог Поль Отле. Наприкінці XIX — початку XX століття запропонував доповнити науку, яка володіла науково-технічною інформацією, і традиційне бібліотекознавство абсолютно новим методом, який він назвав "Документацією"[1].

Для забезпечення проблем доступу до інформації, людство створило бібліотеки як універсальну систему зберігання "знань", їхньої систематизації та каталогізації.

Існуючу проблему відбору інформації вже давно намагаються вирішити шляхом створення універсальних або спеціалізованих інформаційно-пошукових систем. У результаті розвитку технологій пошуку в порівнянні з методиками роботи із семантичною інформацією утворився помітний розрив між технікою роботи з даними (пошуком) і здатністю працювати зі змістом, закладеним у цих даних.

Аналіз останніх досліджень і публікацій. Розроблення систем інформаційного забезпечення різних аспектів наукової діяльності на базі нових інтернет-технологій, присвячені публікації В. Серебрякова, А. Бездушного, С. Мальцевої та ін. Методологія автоматизації процесів опрацювання текстової інформації представлена в роботах Дж. Солтона, Г. Белонова та ін.

Метою статті є дослідження процесу створення інформаційно-пошукових систем, які здатні в автоматизованому режимі виділяти метадані із електронних документів, що уможливлює отримання на основі цих даних нову цінну інформацію і знання.

Виклад основного матеріалу. Сучасні тенденції розвитку бібліотечно-інформаційних технологій та бібліотечно-інформаційної сфери в цілому логічно пов'язані із загальними сучасними тенденціями розвитку суспільства, насамперед з постійно зростаючим рівнем високих інформаційних та комп'ютерно-комунікаційних технологій.

Першим зародженням інформаційного простору цивілізації стали найбільші громадські бібліотеки (Бібліотека Британського музею, Національна бібліотека в Парижі, Бібліотека Конгресу, Російська державна бібліотека та ін.).

Тривалий час одним з потужних інструментів пошуку інформації в книжкових сховищах був безпосередній доступ читачів до книг. Нагромадження їх призвело до парадоксального результату, пов'язаного з відділенням книжкових сховищ від широкого кола читачів. Універсальним інструментом пошуку знань, що базується на прямому доступі до інформації, могли користуватись не всі. Більшість була змушена задовольнятися тільки пошуком в каталозі, який не міг задовольнити інформаційні потреби, які виникали. Для вирішення проблеми доступу читачів до інформації були зроблені спроби класифікації та систематизації інформації — стали створюватися спеціалізовані книжкові зали, куди джерела інформації відбиралися, виходячи з певних критеріїв.

У міру нагромадження книг та інформації, що містилася у них, можливості традиційних методів пошуку: з використанням алфавітного каталогу (пошук книги за відомим іменем автора) і систематичного каталогу (пошук книги або класу книг з певного предмета) — перестали задовольняти читачів, насамперед науковців, інформаційні потреби яких у процесі наукового пошуку характеризувалися невисокою чіткістю усвідомлення і вираження [2].

Сучасні інформаційні технології надають досліднику потужний апарат для "маніпулювання даними". Переведені в електронну форму, вони набувають нову якість, забезпечуючи ширше розповсюдження та ефективне використання. На перший погляд, може скластися враження, що розвиток інформаційних технологій уже сам по собі здатний вивести роботу з науковою інформацією на якісно новий рівень, але, на жаль, це зовсім не так, оскільки інформаційні технології поки не можуть надати адекватний апарат для оперування "інформацією" та інформаційними ресурсами.

У працях К. Муерса і Дж. Солтона вживається термін Information Retrieval System (IRS). У 1950—1970-ті роки англомовний термін Information Retrieval (IR) перекладався, як "інформаційний пошук" і відповідно системи цього класу називали інформаційно-пошуковими. У цих системах використовувалися ручні процедури індексування документів і створення тезаурусів. Але, що надзвичайно важливо, вони призначалися для виділення інформації з різних документів. "Виділення" — це точніше значення слова "retrieval". Зараз в енциклопедіях поняття інформаційні ресурси (IP) визначається як наука пошуку інформації в документах і пошуку власне документів та їхнього опису в базах даних (у тому числі мережевих) [3, 4].

Найрадикальніший етап "інформаційної революції" розпочався в останнє десятиріччя минулого століття. Він був пов'язаний із створенням WWW-сервісу мережі Інтернет, а також з масовим розповсюдженням потужних персо-

нальних комп'ютерів, завдяки чому користувачі отримали доступ до ресурсів цього сервера. Саме WWW-сервіс, що відрізняється від друкованих видань оперативністю розміщення та передаванням інформації практично будь-якого характеру, а від класичних електронних ЗМІ — можливістю передаванням друкованого тексту, що формує все більш реальну перспективу створення єдиного інформаційного простору.

На сьогодні Інтернет — головне джерело електронних документів. Говорячи про засоби інформаційного пошуку в його мережі, зазвичай мають на увазі пошукові системи з їхньою можливістю пошуку інформації по всьому Інтернету (принаймні, за всіма www-сторінками). А вони відомі всім користувачам Інтернету: Google, Yahoo, Yandex та ін. Однак для знаходження документів, користувачі часто звертаються до тематичних каталогів інтернет-ресурсів, що є структурованими наборами посилань на документи відповідної тематики [5].

Інформаційно-пошукова система (ІПС) — це сукупність довідково-інформаційного фонду і технічних засобів інформаційного пошуку в ньому. У свою чергу, довідково-інформаційний фонд (ДІФ) — це сукупність інформаційних масивів (упорядкованих сукупностей документів, фактів або відомостей про них) і пов'язаного з ними довідково-пошукового апарату (тобто даних про адреси зберігання документів з певними пошуковими образами документа). Пошуковий образ документа — це текст, що складається з лексичних одиниць інформаційно-пошукової мови (тобто спеціального формалізованої штучної мови), що виражає основний смисловий зміст документа і призначений для реалізації інформаційного пошуку. Процес вираження змісту документа на інформаційно-пошуковій мові називається індексуванням [6].

Під документом у цьому контексті, як правило, розуміють не тільки короткий виклад того, про що описує документ, а його бібліографічний опис: назва, прізвища його авторів, вихідні дані. Сукупність виділених у процесі індексації характеристик документа разом з формальним описом структури цих характеристик зазвичай називають метаданими.

Структурування метаданих спрощує пошук документів, оскільки одне і те саме слово може входити до списку авторів документа, в його назву, анотацію або навіть у вихідні дані. Ці випадки можуть бути розмежовані — саме завдяки структуруванню метаданих.

Документ стає доступним для пошуку за допомогою ІПС, якщо його метаописання (тобто сукупність метаданих) потрапляє в довідково-інформаційний фонд цієї системи. При складанні тематичних каталогів інтернет-ресурсів часто використовуються пошукові роботи, які збирають дані про документи лише з сайтів відповідної тематики. Деякі спеціалізовані інформаційно-пошукові системи створюються виключно ручним способом, при цьому розмір їхніх пошукових масивів може бути вельми значний. Наприклад, одна з найбільших російських наукових електронних бібліотек elibrary.ru [7] увібрала (станом на квітень 2013 р.) реферати та повні тексти понад 15 млн наукових статей і публікацій. База даних наукових публікацій Web of Science — одна з найбільших — [8] містить (станом на квітень 2013 р.) понад 50 млн записів.

Довідково-інформаційні фонди більшості інформаційно-пошукових систем, що працюють з електронними документами, поповнюються не ручним способом, а за допомогою тих чи інших програм, що автоматизують пошук та індексацію документів.

Звичайною практикою універсальних пошукових систем є представлення пошукового образу документа у вигляді неструктурованого набору ключових слів — інформативних слів, приведених до стандартної лексикографічної форми. Інформативними словами називаються слова, словосполучення чи спеціальні позначення у тексті документа (або запиту), які виражають поняття, істотні для передавання змісту документа. Конкретні критерії включення слова або словосполучення до безлічі інформативних слів залежать від виду ІПС. Так, в універсальних пошукових системах як інформативні розглядаються практично всі слова, включаючи службові. У спеціалізованих інформаційно-пошукових системах, для яких набір ключових слів — один з компонентів структури метаданих документа, навпаки, безліч інформативних слів зазвичай будується на основі предметного покажчика відповідної предметної галузі, тоді як слова, що відносяться до "загальновживаної" лексики, до інформативних не включаються [9].

Зважаючи на абсолютно очевидні переваги структурованого опису документа перед неструктурованим, організації, що намагаються виступати "законодавцем" у мережі Інтернет, насамперед консорціум W3C, який неодноразово робив спроби надати засновникам інтернет-документів можливість явно зазначати значення основних елементів метаданих документа, що уможливило б значне підвищення ефективності функціонування пошукових роботів. У специфікації мови гіпертекстової розмітки документів HTML чітко прописано, що у кожного HTML-документа має бути лише один елемент TITLE ("назва") у полі HEAD ("заголовок"). Більше того, в описі мови HTML з'явився елемент META, який призначений для запису парних елементів виду NAME: CONTENT ("назва: значення"), що описують властивості цього документа: прізвище автора, список ключових слів тощо.

Однак специфікація мови HTML не передбачувала конкретних назв для позначення елементів, що містять інформацію про прізвище автора, ключові слова та ін. Найвідомішим підходом до її вирішення став запропонований на семінарі у м. Дублін (штат Огайо, США) базовий набір з 15 полів метаданих, призначений для опису ресурсів, що публікуються в Інтернеті, куди увійшли такі загальні властивості документів, як назва, дата публікації, автор, видавець, власник. Таким чином, у будь-якому документі має існувати ядро метаданих, спосіб інтерпретування яких заздалегідь відомий. Ці пропозиції були опубліковані під робочою назвою метаданих Дублінського ядра, що згодом стали фундаментом проекту Dublin Core Metadata Initiative [10].

Названі ідеї отримали розвиток у проекті Semantic Web, суть якого полягає у створенні мережі документів, які уміщують метадані "вихідних" документів мережі Інтернет та існуючих паралельно з ними. Ця "паралельна" мережа призначена спеціально для побудови пошуковими роботами (та іншими інтелектуальними агентами) однозначних логічних висновків про властивості "вихідних" документів. Основні принципи створення Semantic Web базуються на використанні, по-перше, універсальних ідентифікаторів ресурсів (URI) за допомогою розширення цього поняття на об'єкти, недоступні для скачування з Інтернету, а по-друге, онтологій (тобто формальних моделей описання тих чи інших предметних сфер) та мови опису метаданих.

На жаль, жоден з перерахованих підходів не став по-справжньому широко розповсюдженим. У цьому можна переконатися, переглянувши довільний набір інтернет-документів. Майже в більшості з них відсутні META елементи, що містять прізвища авторів, список ключових слів тощо. Причини ситуації, що склалися, широко досліджуються, але безсумнівно, до основних належить "людський фактор".

По-перше, зважаючи на широку розповсюдженість інтернет-технологій теоретична підготовка багатьох засновників інтернет-ресурсів залишає бажати кращого, вони часто просто не задумуються про призначення елемента META в мові HTML. По-друге, вказівка значень метаданих — процес

досить трудомісткий, тому навіть ті створювачі ресурсів, які знають про технології метаданих, не завжди вважають за потрібне витрачати час і зусилля на роботу з ними, тим паче що розробники універсальних пошукових систем, виходячи з описаної ситуації, не надто покладаються на можливість автоматичного отримання структурованого пошукового образу індексованого документа, тому що відсоток документів, детально описаних засновниками, дуже невеликий.

У дещо кращому становищі перебувають засновники тематичних каталогів інтернет-ресурсів, оскільки кількість організацій, що працюють у тій чи іншій сфері людської діяльності, а також веб-сайтів, що публікують дійсно цінну і нову інформацію відповідної тематики, як правило, досить невелика. Важливо відзначити, що реальні технології створення переважної більшості сайтів такі, що однорідні документи з одного сайту мають практично однакову HTML-розмітку. При цьому неважливо, чи генеруються документи динамічно або ж вони створюються ручним способом за допомогою копіювання вже наявного документа з наступною заміною тексту (що також зберігає розмітку). Ця обставина дає змогу автоматизувати процес індексації метаданих електронного документа за допомогою зазначення шаблону документів того чи іншого сайту, тобто явної вказівки команд (тегів) мови HTML, охоплюючи основні характеристики документа: автори, назва, ключові слова, анотація, коди того чи іншого класифікатора тощо.

Список використаної літератури

1. *Пилко И. С.* Информационные и библиотечные технологии : учеб. пособие / И. С. Пилко. — СПб. : Профессия, 2006. — 342 с.
2. *Полл Р.* Измерение качества работы: Международное руководство по измерению эффективности работы университетских и других научных библиотек / Р. Полл, П. Бокхорст ; пер. с англ. Н. В. Соколовой ; под ред. О. Ю. Устинова. — М. : Логос, 2002. — 152 с.
3. *Попов В. Д.* Информационные процессы в обществе и модели управления ими / В. Д. Попов // Управление общественными отношениями : учебник / [Комаровский В. С. и др.]. — М., 2003. — С. 34—37.
4. *Попов И. И.* Мировые информационные ресурсы и сети: учеб. пособие / И. И. Попов, П. Б. Храмцов. — М. : Изд-во РЭА им. Г. В. Плеханова, 1999. — 145 с.
5. *Сборник* основных российских стандартов по библиотечно-информационной деятельности / сост.: Т. В. Захарчук, О. М. Зусьман. — СПб. : Профессия, 2005. — 547 с.
6. *Сладкова О. Б.* Категория времени в социальной технологии мониторинга / О. Б. Сладкова // Научно-техническая информация. Серия 1. — 2000. — № 3. — С. 1—6.
7. *Elibrary.ru* Научная электронная библиотека. — Mode of access: http://elibra-ry.ru/defaultx.asp. — Title from the screen.
8. *Web* of science. — Mode of access: http://wokinfo.com/products_tools/multidisciplinary/webofscience/. — Title from the screen.
9. *White W G.* The core business web : a guide to key information resources / Gary W White. — Routledge : Barnes & Noble, 2013. — 340 p.
10. *Wroblewski L. Site-seeing: a visual approach to web usability / Luke Wroblewski. — New York : Wiley, 2002. — 364 p.*

В данной статье исследуются различные вопросы разработки и создания информационно-поисковых систем библиотеки, которые способны в автоматизированном режиме получать данные из электронных документов с целью внедрения этих документов в научно-информационный процесс.

This article investigates various issues of development and creation of information retrieval systems library that can out automatically receive data from electronic documents for the implementation of these instruments in scientific and information process.

Надійшла в редакцію 30 квітня 2013 року