УДК 025.4.03:025.5:004.7

Олег Сербін,
завідувач відділу систематизації НБУВ,
канд. іст. наук

ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ ПОШУКОВИХ ІНСТРУМЕНТІВ У КОНТЕКСТІ РОЗВИТКУ ІНДЕКСУВАННЯ ІНФОРМАЦІЙНИХ РЕСУРСІВ

У публікації представлено аналітичний огляд процесу індексування інформації в інтерактивному просторі. Наведено основні типологічні особливості пошукових інструментів та розкрито суть інформаційного пошуку крізь призму структурних особливостей пошукових інструментів в Інтернет. Доведено, що індексування інформації є основною і вагомою щодо ефективності складовою пошукових інструментів.

Ключові слова: індексування, інформація, інформаційний пошук, інформаційно-пошукова система, каталог, мета-система, інформаційно-пошукова мова, читач-користувач.

Великий обсяг інформації не є запорукою інформативності. Адже з безлічі інформаційних даних важливими є ті, які відповідають першочерговим потребам читачів-користувачів. У цьому сенсі головною є можливість виокремлення потрібної інформації з розмаїття її множинності. Саме індексування, на нашу думку, є тим основним процесом, що забезпечує певне виокремлення одиничного. Це стосується як традиційних форм роботи з інформаційними даними, так и новітніх – інтерактивних, які, правду кажучи, з кожним днем дедалі більше і більше стають традиційними, з огляду на поширення та масовість їх використання. Тому огляд основних аспектів індексування в інтерактивному просторі та інструментів пошуку, визначальною частиною яких і є індексування, є актуальним щодо аналізу та розкриття цього питання. Об'єктом цього дослідження є індексування інформації як ефективної складової пошукових інструментів, а предметом – здійснення процесу індексування в межах різних пошукових інструментів. Метою цього дослідження є аналіз та визначення основних аспектів здійснення процесу індексування, враховуючих типологічні особливості пошукових інструментів.

У першу чергу треба зазначити, що пошукові інструменти – це особливе програмне забезпечення, головна мета якого – забезпечити найбільш оптимальний і якісний пошук інформації для читачів-користувачів Інтернету. Пошукові інструменти розміщуються на спеціальних веб-серверах, кожен з яких виконує певну функцію: аналіз веб-сторінок і занесення результатів аналізу на той чи інший рівень бази даних пошукового сервера; пошук інформації за запитом користувача; забезпечення зручного інтерфейсу [2] для пошуку інформації та перегляду результату пошуку користувачем.

У даному випадку процес пошуку забезпечується за допомогою здійснення простих операцій. Спочатку користувач вводить ключову фразу і активізує пошук, тим самим отримує добірку документів за сформульованим (заданим) запитом. Цей список документів ранжирується за певними критеріями так, щоб вгорі списку опинилися ті документи, які найбільше відповідають запиту користувача. Кожен з пошукових інструментів використовує різні критерії ранжирування документів, як під час аналізу результатів пошуку, так і під час формування індексу [3] (наповнення індексної бази даних веб-сторінок).

Таким чином, якщо вказати в рядку пошуку для кожного пошукового інструменту однакової конструкції запит, можна отримати різні результати пошуку. Для користувача має велике значення, які документи будуть у перших двох-трьох десятках документів за результатами пошуку і наскільки ці документи відповідають очікуванням користувача. Це, у свою чергу, зумовлено якістю інформаційного пошуку як такого.

Треба зазначити, що інформаційний пошук – процес пошуку в деякій множині текстів (документів) усіх таких, які присвячені зазначеній у запиті темі (предмету) [1] або містять потрібні споживачеві факти, відомості. Інформаційний пошук здійснюється за допомогою інформаційно-пошукової системи і виконується вручну або з використанням засобів механізації або автоматизації. Неодмінним учасником інформаційного пошуку є людина – читач-користувач. Залежно від характеру інформації, яка міститься у виданих інформаційно-пошуковою системою текстах, інформаційний пошук може бути документальним, у тому числі бібліографічним, і фактографічним. Інформаційний пошук потрібно

[2] Інтерфейс пошукового інструменту представлений у вигляді сторінки з гіперпосиланнями, рядком подачі запиту (рядком пошуку) та інструментами активізації запиту [7].

[3] Індекс пошукової системи – це інформаційна база, яка містить результат аналізу веб-сторінок, складена за певними правилами [2].

відрізняти від логічної переробки інформації, без якої неможлива безпосередня видача людині відповідей на поставлені нею запитання. Під час інформаційного пошуку можуть бути знайдені такі і лише такі факти або відомості, які були введені в ІПС. Перед введенням в ІПС тексту (документа) визначається його основний смисловий зміст (тема або предмет), який потім переводиться і записується на одну з інформаційно-пошукових мов. Цей запис називається пошуковим образом тексту [4]. Так само роблять і коли в ІПС вводять певним чином записані факти, відомості. Запит, що надійшов, також переводиться на інформаційно-пошукову мову, утворюючи пошуковий припис. Оскільки пошукові образи текстів і пошукові розпорядження записані однією і тією самою мовою, вирази якою допускають тільки одне тлумачення, то можливо порівнювати їх формально, не вникаючи в сенс. Для цього задаються певні правила (критерії відповідності), що встановлюють при якій мірі формального збігу пошукового образу з пошуковим розпорядженням [5] текст слід вважати таким, що відповідає на інформаційний запит [6] і є предметом видачі [7].

Технічна ефективність інформаційного пошуку характеризується двома відносними показниками – коефіцієнтом точності (відношенням числа текстів, що відповідають на інформаційний запит, до загальної кількості текстів у даній видачі) і коефіцієнтом повноти (відношенням числа текстів, що відповідають на інформаційний запит, до загальної кількості таких текстів, що містяться в даній ІПС). Необхідні значення цих показників залежать від специфіки інформаційних потреб.

З іншого боку, інформаційний пошук може бути двох типів – вибіркове (або адресне) поширення інформації та ретроспективний пошук. При виборчому поширенні інформації інформаційний пошук проводиться за постійними запитами деякої кількості споживачів (абонентів), здійснюється періодично (зазвичай один раз на тиждень або на два тижні)

---

[4] Текст, що складається з лексичних одиниць інформаційно-пошукової мови, що виражає зміст документа або інформаційного запиту і призначений для реалізації інформаційного пошуку [5].

[5] Таким чином, пошуковий припис – це виражене в термінах формалізованої інформаційно-пошукової мови смисловий зміст інформаційного запиту [6].

[6] Ідеться про ключове слово або фразу, яку вводить користувач у рядок пошуку. Для формування різних запитів використовуються спеціальні символи («», ~), математичні символи (*, +,?) [8].

[7] Або сторінка результатів пошуку (англ. Search engine results page, SERP) – веб-сторінка, що генерується пошуковою системою у відповідь на пошуковий запит користувача [9].

і виконується лише в масиві текстів, що надійшли в ІПС за цей період часу. Між ІПС і читачами-користувачами встановлюється ефективно діючий зворотний зв'язок (читач-користувач повідомляє, у якій мірі цей текст відповідає запиту і чи потрібна йому копія повного тексту, про ступінь відповідності цього тексту його інформаційній потребі), який дає змогу уточнювати потреби абонентів, своєчасно реагувати на зміни цих потреб і оптимізувати роботу системи.

При ретроспективному пошуку ІПС відшукує необхідну інформацію у всьому накопиченому масиві текстів за разовими запитами.

Як бачимо, суть та значення інформаційного пошуку виправдано вагомі, але не менше важливим є те, яким саме чином здійснювати цей пошук, засобами якого інструментарію. Безумовно, можна шукати потрібні джерела інформації вручну, дізнаватися адреси із спеціалізованих журналів з інформатики та Інтернету, використовувати спеціальні паперові довідники з класифікованими за категоріями адресами. Однак для такого мінливого простору як Інтернет необхідно навчитися користуватися спеціальними інструментами, мета яких – збирати дані про інформаційні ресурси та надавати користувачам послуги швидкого пошуку. У загальному випадку, можна виділити такі пошукові інструменти для всесвітньої павутини: каталоги, пошукові системи, автономні пошукові агенти.

Розкриваючи значення першого з перерахованих інструментів, зазначимо, що каталог являє собою по суті пошукову систему з класифікованих [8] за темами списком анотацій з посиланнями на веб-ресурси.

Пошук у каталозі проводиться за допомогою послідовного уточнення тем. Тим не менш, каталоги підтримують можливість швидкого пошуку певної категорії або сторінки за ключовими словами за допомогою локальної пошукової машини. База даних посилань (індекс) каталогу зазвичай має обмежений обсяг і заповнюється, як вже зазначалось, вручну персоналом каталогу. Деякі каталоги використовують автоматичне оновлення індексу. Фактично результат пошуку в каталозі представляється у вигляді списку, що складається з короткого опису (анотації) документів з гіпертекстовим посиланням на першоджерело.

Цікавим є те, що засновники та розробники перших пошукових систем скористалися методом пошуку книг у бібліотеках. Вони створили тематичні каталоги, у категоріях яких і розташовувалися потрібні сайти. Читач-користувач заходив у каталог, вибирав потрібну рубрику та отримував кілька сайтів з тематики, що до неї належать. На початку,

[8] Процес класифікації здійснюється зазвичай людиною *(прим. О. Сербіна)*.

коли сайтів було не багато, все було прекрасно, а потім стало дедалі складніше відшукати потрібний ресурс. Рубрик ставало щораз більше і більше, вкладеність категорій дедалі зростала. Нарешті, доводилося продиратися шлях у безліч підкатегорій, що призводив до неповноти видачі або ж узагалі до інформаційного мовчання. Пізніше всі пошукові системи трансформувались у пошукові покажчики [9]. На відміну від каталогів, пошукові покажчики формують зв'язки «запит-відповідь», і до «відповідей» могли приписуватися декілька різних ресурсів. Але згодом з'ясувалося, що це важко не тільки для людей, а й для пошукових машин, тому що на будь-який поширений запит отримувалися сотні тисяч відповідей, у яких просто неможливо було розібратися. І саме тому можливість вибору пошукувачем із цього розмаїття інформації ряду релевантних и влучних посилань стало визначальним щодо подальшого широко застосування ІПС.

Більше того, на відміну від ІПС, застосовуючи каталоги, не використовуються спеціальні інструменти, щоб індексувати сторінку або веб-сайт. Для реєстрації в каталозі необхідно знайти розділ, у який потрібно помістити певну сторінку, надіслати короткий опис сайту і список ключових слів для пошуку вашої сторінки в каталозі. Потім ця інформація проглядається й оцінюється фахівцями, які вирішують, чи варто включати дану інформацію в каталог чи ні.

У разі ж використання пошукових систем для індексування сайтів застосовуються так звані пошукові роботи – невелика програма, яка «ходить» по посиланнях на сайти та індексує (збирає і запам'ятовує) знайдену на шляху інформацію. Весь процес індексування здійснюється таким чином: від читача-користувача пошукова система отримує точну адресу сторінки, яку потрібно зареєструвати. Пошукова система перевіряє, чи існує за цією адресою сторінка, і якщо так, то сторінка вноситься в графік відвідування.

Таким чином, пошукова інформаційна система – це організована сукупність програмно-технічних та інших допоміжних засобів, технологічних процесів і функціонально-певних груп працівників, які забезпечують збір, представлення й накопичення інформаційних ресурсів у певній предметній області, пошук і видачу відомостей, необхідних для задоволення інформаційних потреб користувачів. Вона забезпечує пошук і відбір необхідних даних у спеціальній базі з описами джерел

[9] Пошукова система, у якій інформаційний масив динамічний, тобто запит читача-користувача аналізується, і у відповідь на різні запити він отримує різні відповіді *(прим. О. Сербіна)*.

інформації (індексу) на основі інформаційно-пошукової мови і відповідних правил пошуку.

Слід зазначити, що ІПМ – це знакова система, призначена для опису (шляхом індексування) основного смислового змісту текстів (документів) або їх частин, а також для вираження смислового змісту інформаційних запитів з метою реалізації інформаційного пошуку. Будь-яка абстрактна ІПМ складається з алфавіту (списку елементарних символів), правил освіти і правил інтерпретації. Правила освіти встановлюють, які комбінації елементарних символів допускаються під час побудови слів і виразів, а правила інтерпретації – як належить розуміти ці слова й вирази. ІПМ повинна розташовувати лексико-граматичні засоби, необхідні для вираження основного смислового змісту будь-якого тексту і сенсу будь-якого інформаційного запиту з даної галузі або предмета, бути недвозначною (допускати одне тлумачення кожного запису), зручною для алгоритмічного зіставлення й ототожнення (повного або часткового) записів основного смислового змісту текстів і смислового змісту інформаційних запитів. Під час розробки конкретної ІПМ враховуються специфіка галузі або предмета, для якого ця мова створюється, особливості текстів, утворюється пошуковий масив, характер інформаційних потреб, для задоволення яких створюється дана інформаційно-пошукова система. У більшості ІПМ основний словниковий склад задається його перерахуванням і являє собою фрагмент лексики тієї чи іншої природної мови. Відібрані з природної мови слова і словосполучення в сукупності утворюють основний словниковий склад, служать як би алфавітом даної ІПМ. Правила освіти в такій ІПМ виконують функцію синтаксису. У деяких ІПМ основний словниковий склад задається (повністю або частково) методом породження, який полягає в тому, що для таких ІПМ правила освіти встановлюють, як з даного алфавіту будувати слова ІПМ, а з цих слів – вирази (фрази) і які з них будуть правильно побудованими. Варто зауважити, що із середини XX ст. як ІПМ широко використовуються бібліотечно-бібліографічні класифікації (класифікаційні ІПМ), переліки предметних рубрик (предметизаційна ІПМ) та тезаурус (дескрипторна ІПМ).

Використання адекватної ІПМ спрямовано на виконання головного завдання будь-якої ІПС – це пошук інформації релевантної інформаційним потребам користувача. Дуже важливо в результаті проведеного пошуку нічого не втратити, тобто знайти всі документи, пов’язані з запитом, і не знайти нічого зайвого. Тому вводиться якісна характеристика процедури пошуку – релевантність – відповідність результатів пошуку сформульо-

ваному запиту. Релевантність, поряд з повнотою бази, обліком морфології мови, є основним критерієм якості роботи пошукової системи.

На сьогодні більшість пошукових систем перейшло на чотирьохетапну структуру роботи. Першим і найбільш важливим етапом у даному випадку є індексування інформації, адже відповідно до того, яким саме чином буде індексована інформація, покладено і хід усіх подальших операцій щодо роботи з інформаційними даними. Отже, спочатку пошукова система індексує інформацію і заносить її у базу даних, потім, з урахуванням морфології, розрізає всі слова сторінок на складові. Індексування – процес вираження головного предмета або теми тексту якого-небудь документа в термінах інформаційно-пошукової мови. Застосовується для полегшення пошуку необхідного тексту серед безлічі інших. Проводиться індексування як цілого документа, так і його частини. Для індексування нерідко використовуються заголовки текстів. При індексуванні опускаються супутні предмети або теми. Це слугує причиною того, що під час пошуку не знайденими залишаються тексти, для яких предмет або тема інформаційного запиту є не головною, а супутньою.

Узагалі розрізняють два основні типи індексування – класифікаційне і координатне [3, с. 5]. При класифікаційному індексуванні, або класифікуванні, тексти залежно від їх змісту включаються до відповідного одного або декількох класів, у якому збираються всі тексти, що мають в основному однаковий смисловий зміст. Кожному такому тексту присвоюється індекс цього класу, який далі слугує його пошуковим образом.

Координатне індексування характеризується тим, що основний смисловий зміст тексту виражається переліком повнозначних слів, які обирають або із самого тексту, або його заголовку, або зі спеціального нормативного словника. У першому випадку такі лексичні одиниці називаються ключовими словами, а в другому – дескрипторами. Кожне ключове слово або дескриптор означає клас, до якого потенційно входять всі тексти, де у вираження основного смислового змісту входить це слово. Логічне створення класів, які позначені всіма словами, виражають в сукупності основний смисловий зміст тексту, якби утворює деякий складний клас. Побудований таким способом складний клас позначається переліком ключових слів або дескрипторів, і цей перелік слугує пошуковим образом даного тексту або вираженням на інформаційно-пошуковій мові смислового змісту запиту. Таким чином, при координатному індексуванні смисловий зміст тексту виражається як би зазначенням його координат у деякому n-мірному смисловому просторі. Різновидом координатного індексування є пермутаційне [10]

або циклічне індексування, що базується на використанні ключових слів заголовку тексту і полягає в тому, що всі ключові слова заголовка разом з контекстом по черзі виводяться в пошукову колонку, де ключові слова наводяться в алфавітному порядку. Координатне індексування не створює ніяких труднощів при пошуку текстів по будь-якому, заздалегідь не передбаченому поєднанню ознак і це його основна перевага над класифікаційним індексуванням.

Слід також зауважити, що особливим типом індексування є розкриття смислового змісту тексту через приведення разом з ним бібліографії, тобто імена авторів і бібліографічні описи їх робіт, на які посилається автор даного тексту. Таке індексування служить основою для складання покажчиків цитованої літератури – вельми ефективного інструменту не тільки для пошуку документів [4, с. 14], а й для вирішення інших завдань (наукознавчих, прогностичних і т. д.). Це суттєво полегшує роботу и підвищує професіоналізм читачів-користувачів у межах проведення їх науково-дослідної роботи зокрема.

І наостанок, варто акцентувати увагу на інструментах пошуку, які не формують власний індекс, але вміють використовувати можливості інших пошукових систем. Ідеться про мета-пошукові системи (пошукові служби) – системи, здатні послати запити користувача одночасно декільком пошуковим серверам, потім об'єднати отримані результати та представити їх користувачеві у вигляді документа з посиланнями.

У даному випадку можна говорити не просто про інформаційний пошук, а про мета-пошук, який може зробити сайт більш помітним для користувачів. Його перевага в тому, що в рамках одного пошуку використовуються кращі з боку багатьох провідних пошукових систем. Особливість у тому, що не можливо просто включити сайт в індекс мета-пошукових систем, тому що в них просто немає своїх баз даних. Спочатку потрібно правильно зареєструватися у звичайних пошукових системах, а мета-пошукувачі використовують їх результати пошуку під час обробки своїх запитів.

До недавнього часу кожна окрема традиційна пошукова система індексувала незрівнянно менший обсяг даних, ніж той, який враховувався при здійсненні мета-пошуку. Зараз, з появою пошукувачів із глибокою індексацією, ситуація поступово змінюється. Але все ж використання мета-пошукових систем істотно розширює зону пошуку, оскільки вони опитують безліч баз даних.

Отже, підсумовуючи вищепроведений аналітичний огляд, можна констатувати, що пошукові інструменти – це особливе програмне забезпе-

чення, головна мета якого забезпечити найбільш оптимальний і якісний пошук інформації для читачів-користувачів Інтернету. У результаті типологічного аналізу встановлено, що існують такі пошукові інструменти: каталоги, пошукові системи, мета-пошукові системи.

Визначено, що існує два основні типи індексування – класифікаційне і координатне. У першому випадку, тексти залежно від їх змісту включаються до відповідного одного або кількох класів, у якому (яких) збираються всі тексти, що мають в основному однаковий смисловий зміст. В іншому випадку, основний смисловий зміст тексту виражається переліком повнозначних слів, які обираються або із самого тексту чи його заголовку (ключові слова), або зі спеціального нормативного словника (дескриптора).

Таким чином, проаналізувавши основні аспекти процесу індексування в інтерактивному просторі, можна зробити висновок про першочерговість зазначеного процесу як елементу пошуку інформації. Сам пошук інформації можна здійснювати внаслідок застосування потенціалу й можливості пошукових інструментів, що так чи інакше доповнюють або взаємозаміняють один одного і тим самим забезпечується компетентна наповненість результату пошуку, його релевантність і, як наслідок, задоволення інформаційних потреб читача-користувача.

Список використаних джерел

1. Возможности применения информационных и коммуникационных технологий в открытом образовании [Электронный ресурс]. – Режим доступа: URL: http://www.ido.rudn.ru/Open/ikt/3.htm. – Дата доступа: 29.03.2012. – Загл. с экрана.

2. Индекс поисковой системы. Что такое «индекс поисковой системы»? [Электронный ресурс]. – Режим доступа: URL: http://westseo.ru/article/search-engine-index/. – Дата доступа: 29.03. 2012. – Загл. с экрана.

3. Индексирование документов. Общие требования к систематизации и предметизации : ГОСТ 7.59–2003. – Изд. офиц. – М. : Изд-во стандартов, 2003. – 8 с.

4. Индексирование документов. Общие требования к систематизации и предметизации: инструктивно-методические указания / сост. Сукиасян Э. Р. – М., 1991. – 61 с.

5. По характеру использования информации [Электронный ресурс]. – Режим доступа: URL: http://www.itstan.ru/it-i-is/po-harakteru-ispolzovanija-informacii.html. – Дата доступа: 29.03.2012. – Загл. с экрана.

6. Поисковое предписание [Электронный ресурс]. – Режим доступа: URL: http://www.ngpedia.ru/id304924 p1.html. – Дата доступа: 29.03.2012. – Загл. с экрана.

7. Сетевые службы, клиенты, серверы, ресурсы. Файлове архивы. Поиск информации в Интернете [Электронный ресурс]. – Режим доступа: URL: www.dist.sch871.edusite.ru/DswMedia/faylovyiearxivyi.docx. – Дата доступа: 29.03.2012. – Загл. с экрана.

8. Технология поиска информации в Интернет. Виды поисковых инструментов [Электронный ресурс]. – Режим доступа: URL: http://www.seonews.ru/masterclasses/detail/29812.php. – Дата доступа: 29.03.2012. – Загл. с экрана.

9. Search engine results page [Electronic resource]. – Mode of access: URL: http://www.ajazi.com/search-engine-results-page.cfm. – Date of access: 2 April 2012. – Title from the screen.

10. What is a permuted index? [Electronic resource]. – Mode of access: URL: http://cboard.cprogramming.com/cplusplus-programming/15192-what-permuted-index.html. – Date of access: 2 April 2012. – Title from the screen.