

УДК 004.9

Ю.Е. Парфьонов, О.І. Морозов

*Харківський національний економічний університет імені Семена Кузнеця, Харків*

## АНАЛІЗ МЕТОДІВ ІНФОРМАЦІЙНОГО ПОШУКУ В МЕРЕЖІ ІНТЕРНЕТ

*У статті розглянуті основні методи інформаційного пошуку в мережі Інтернет. Проведено огляд досліджень щодо інформаційного пошуку та визначені основні його завдання. Розглянуті переваги та недоліки пошукових машин і наведені показники ефективності пошуку інформації в Інтернеті. Зроблені висновки про необхідність використання певних пошукових засобів для отримання якісних результатів пошуку.*

**Ключові слова:** інформаційний пошук, пошукова машина, Інтернет, веб-сторінка, релевантність.

### Вступ

Пошук інформації є одним з найбільш поширених і водночас найбільш складних завдань, з якими доводиться стикатися в мережі Інтернет будь-якому користувачеві. Однак, якщо для звичайного користувача мережевого співтовариства знання методів ефективного інформаційного пошуку є бажаним, але далеко не обов'язковим, то для професіоналів інформаційної діяльності вміння швидко орієнтуватися в ресурсах Інтернет і знаходити необхідні джерела відноситься до числа базових кваліфікаційних навичок.

Причина складнощів, що виникають при інформаційному пошуку в Інтернеті, визначається двома головними факторами. По-перше, число джерел в мережі Інтернет надзвичайно велике. Аналітики підкреслюють, що обсяг інформації, який зберігається в Інтернеті, подвоюється приблизно кожні півтора року. За оцінками IDC, у 2012 році сумарний обсяг контенту Всесвітньої мережі збільшився до 2 500 екзабайт. При цьому в 2006 році в мережі зберігалось всього 161 млрд. Гб даних [1]. По-друге, масив інформації в мережі Інтернет не тільки колосальний за обсягом, але ще і вкрай динамічний. За кожні півхвилини у віртуальному всесвіті з'являється близько сотні нових або змінених документів, десятки переміщуються на нові адреси, а одиниці - назавжди припиняють своє існування. Інтернет ніколи не стоїть на місці, як наша планета, по якій безперервно котиться хвиля ділової активності людства в точній відповідності зі зміною часових поясів [2].

Інтернет призначений для обміну інформацією між людьми. Функції обміну інформацією реалізують служби мережі Інтернет. Причина бурхливого зростання мережі Інтернет полягає не тільки в тому, що її служби пропонують зручні засоби для обміну інформацією та доступу до неї, а й у тому, що в

мережі завжди є необхідна інформація. Служби мережі Інтернет працюють цілодобово і без вихідних.

Величезні інформаційні ресурси стали доступні завдяки Інтернет-службі World Wide Web (WWW). Інформаційні ресурси (веб-сторінки) WWW створюються самими користувачами. Для створення публікації користувачеві потрібен лише комп'ютер, підключений до мережі Інтернет, і мінімум навичок роботи з ним. Опублікований документ стає доступним всім користувачам мережі. Веб-сторінки можуть включати звук, відео, анімацію, що значно підвищує сприйняття інформації користувачем. Веб-сторінка зазвичай включає в себе гіпертекстові посилання (гіперпосилання), що дозволяють відкрити іншу сторінку або переміститися за посиланням на відповідну частину поточної сторінки. Гіпертекст являє собою текст зі вставленими в нього командами розмітки, які посилаються на інші місця цього тексту, інші документи, зображення і т.д. При підготовці гіпертекстових документів для WWW текст розмічається за допомогою мови HTML (HyperText Markup Language – мова розмітки гіпертекстів). Гіпертекстові файли мають розширення \*.htm або \*.html. Як і більшість інших служб Інтернету, служба WWW працює в рамках моделі клієнт-сервер. В якості сервера, як правило, виступає постійно підключений до мережі Інтернет комп'ютер, на якому працює спеціальна програма – веб-сервер. Клієнтом є будь-який комп'ютер, підключений в даний момент до Інтернету, на якому запущена програма перегляду веб-публікацій – браузер. Робота браузера полягає в обміні інформацією з веб-сервером, отриманні необхідних користувачеві документів, обробці отриманої гіпертекстової інформації і відображення документа на екрані [3,4].

### Виклад основного матеріалу

Пошук інформації – одне із завдань, яке найчастіше доводиться вирішувати будь-якому

користувачу Інтернету. Але знайти у великій кількості сайтів і веб-сторінок необхідну і точну інформацію та потрібні ресурси – дуже непросто. Тому треба вміти використовувати різні способи пошуку інформації, правильно формулювати запити й критично оцінювати знайдену інформацію. Завдання інформаційного пошуку є предметом наукових досліджень вже кілька десятиріч. Ще не так давно дослідження в цій області належали до сфери наукових інтересів вузької групи фахівців. Проте бурхливий розвиток Інтернет кардинально змінив ситуацію. Він не тільки привернув увагу до області інформаційного пошуку, але також сильно розширив список розглянутих завдань. Сьогодні дослідження в цій області присвячені не тільки питанням індексування та пошуку в колекції текстових документів, а й моделюванню, завданням класифікації та категоризації документів, архітектурі пошукових систем, питанням візуалізації даних, інтерфейсів користувача і мовам запитів [5].

Всі завдання інформаційного пошуку вирішуються завдяки використанню різноманітних методів. Можна виділити наступні основні методи пошуку інформації в Інтернеті, які, в залежності від цілей і завдань використовуються окремо або в комбінації один з одним: безпосередній пошук з використанням гіпертекстових посилань; пошук з використанням пошукових машин; пошук із застосуванням спеціальних пошукових засобів; пошук по новоствореним ресурсам.

Що стосується безпосереднього пошуку з використанням гіпертекстових посилань, то можна сказати, що всі сайти в просторі WWW фактично виявляються пов'язаними між собою, пошук інформації може бути проведений шляхом послідовного перегляду пов'язаних сторінок за допомогою браузерів.

Хоча цей повністю ручний метод пошуку виглядає повним анахронізмом у мережі Інтернет, що містить більше 60 млн. вузлів, "ручний" перегляд веб-сторінок часто виявляється єдиною можливістю на заключних етапах інформаційного пошуку, коли механічне "копання" поступається місцем більш глибокому аналізу.

Використання каталогів, класифікованих і тематичних списків і всіляких невеликих довідників також належить до цього виду пошуку.

Використання пошукових машин сьогодні є одним з основних методів і фактично єдиним при проведенні попереднього пошуку. Результатом останнього може бути список ресурсів мережі Інтернет, що підлягають детальному розгляду.

Як правило, застосування пошукових машин засноване на використанні ключових слів, які

передаються до пошукових серверів у якості аргументів пошуку.

Пошук із застосуванням спеціальних пошукових засобів – це повністю автоматизований метод, який виявляється досить ефективним для проведення первинного пошуку. Одна з технологій цього методу заснована на застосуванні спеціалізованих програм – спайдерів, які в автоматичному режимі переглядають веб-сторінки та шукають необхідну інформацію.

Фактично це автоматизований варіант перегляду за допомогою гіпертекстових посилань, описаний вище (пошукові машини для побудови своїх індексних таблиць використовують схожі методи).

Немає потреби говорити, що результати автоматичного пошуку обов'язково вимагають подальшої обробки.

Застосування даного методу доцільно, якщо використання пошукових машин не може дати необхідних результатів (наприклад, в силу нестандартності запиту, який не може бути адекватно заданий існуючими засобами пошукових машин). У ряді випадків цей метод може бути дуже ефективним.

Вибір між використанням спайдера або пошукових машин являє собою варіант класичного вибору між застосуванням універсальних або спеціалізованих засобів.

Пошук по новоствореним ресурсам може виявитися необхідним при пошуку найбільш актуальної інформації, проведенні повторних циклів пошуку або для аналізу тенденцій розвитку об'єкта дослідження в динаміці.

Іншою можливою причиною може з'явитися те, що більшість пошукових машин оновлює свої індекси зі значною затримкою, викликані гігантськими обсягами оброблюваних даних, і ця затримка зазвичай тим більше, чим менша популярна тема вас цікавить. Це міркування може виявитися досить істотним при проведенні пошуку в вузькоспеціалізованій предметній області [6].

Засоби пошуку використовуються для того, щоб допомогти людям знайти інформацію, якої вони потребують. Засоби пошуку типу агентів і спайдерів використовуються для збору інформації про документи, які перебувають у мережі Інтернет. Це спеціальні програми, які займаються пошуком сторінок в Інтернеті, отримують гіпертекстові посилання з цих сторінок і автоматично індексують інформацію, яку вони знаходять для побудови бази даних. Кожний пошуковий механізм має власний набір правил, що визначають, як збирати документи. Деякі слідує за кожним посиланням на кожній знайдений сторінці і потім, у свою чергу, досліджують кожне посилання на кожній з нових сторінок, і так далі. Деякі ігнорують посилання, які ведуть до графічних і звукових файлів, файлів мультимедіа; інші ігнорують посилання до ресурсів

типу баз даних; інші проінструковані, що потрібно переглядати насамперед найбільш популярні сторінки.

Агенти – "найінтелектуальніші" з пошукових засобів. Вони можуть робити більше, ніж просто шукати: вони можуть виконувати навіть транзакції від вашого імені. Вже зараз вони можуть шукати сайти специфічної тематики і повертати списки сайтів, відсортованих за їх відвідуваністю. Агенти можуть обробляти вміст документів, знаходити та індексувати інші види ресурсів, не лише сторінки. Вони можуть бути запрограмовані для витягання інформації з вже існуючих баз даних. Незалежно від інформації, яку агенти індексують, вони передають її назад до бази даних пошукового механізму.

Спайдер - це ключовий інструмент для пошуку в Інтернеті. Спайдери повідомляють про зміст знайденого документа, індексують його і добувають підсумкову інформацію. Вони також переглядають заголовки, деякі посилання і відправляють проіндексовану інформацію до бази даних пошукового механізму [7].

Коли хто-небудь хоче знайти інформацію, доступну в Інтернеті, він відвідує Інтернет-сторінку пошукової системи і заповнює форму, що деталізує інформацію, яка йому необхідна і починає пошук. Тут можуть використовуватись ключові слова, дати та інші критерії. Критерії в формі пошуку повинні відповідати критеріям, які використовуються агентами при індексації інформації, яку вони знайшли при переміщенні по мережі Інтернет.

Пошукова машина виводить ранжований перелік веб-сторінок людині, що зробила запит. Різні пошукові механізми також вибирають різні способи показу отриманого списку - деякі відображають лише посилання, інші виводять посилання з першими кількома пропозиціями, що містяться в документі або заголовках документу разом з посиланням.

До переваг пошукових машин слід віднести наступні: мала кількість застарілих посилань в результатах пошуку; набагато більша кількість веб-сайтів, за якими проводиться пошук; більш висока швидкість пошуку; висока релевантність пошуку; наявність додаткових сервісних функцій, що полегшують роботу користувача, наприклад, можливість перекладу тексту документа на іноземну мову, здатність виділяти всі документи з певного сайту, звуження критеріїв у ході пошуку і т.д.

Однак пошукові машини мають обмежену область пошуку. Якщо який-небудь сайт не був внесений в базу даних пошукової системи, він для неї не існує, і його документи в результати пошуку потрапити не можуть. Для того, щоб складений запит на пошук точно відповідав тому, що саме потрібно знайти, потрібно хоча б трохи уявляти, як працює пошукова система, і вміти використовувати

найпростіші логічні оператори та синтаксис пошукових запитів [8].

Пошук з використанням пошукових машин, є найпоширенішим і ефективним методом пошуку чогось конкретного в мережі Інтернет. Знання різних засобів, які допомагають знайти потрібну інформацію дуже важливі, оскільки в Інтернеті опубліковані десятки мільйонів сторінок, то швидко відшукати потрібну інформацію досить складно. Зазвичай при пошуку веб-сторінок з використанням пошукової машини в якості результату видається занадто великий обсяг інформації. Тому чим сильніше ви звужите діапазон пошуку, тим точніше буде результат.

Одними з найбільш важливих показників ефективності систем інформаційного пошуку є семантичні показники. Семантичні показники засновані на оцінці релевантності між знайденими документами і пошуковими запитами.

Релевантність – це об'єктивно-існуюча смислова відповідність між змістом документа і пошуковим запитом.

Об'єктивність оцінки релевантності забезпечується тим, що вона встановлюється експертним шляхом, а не автором пошукового запиту.

Семантичними показниками є повнота видачі, втрата інформації, точність видачі, інформаційний шум.

Введемо наступні позначення:

а – множина релевантних та виданих системою документів;

б – множина не релевантних, але виданих системою документів;

в – множина релевантних, але не виданих системою документів.

Тоді повнота видачі інформації ( буде дорівнювати:

(1)

Точність видачі ( буде дорівнювати:

(2)

Втрата інформації ( буде дорівнювати:

(3)

Інформаційний шум ( буде дорівнювати:

(4)

Іншою групою показників ефективності систем інформаційного пошуку є прагматичні показники. Ці показники можуть визначати тільки в процесі експлуатації інформаційної системи. Прагматичні показники визначають абоненти системи на базі оцінок пертинентности виданих документів.

Пертинентність – це суб'єктивно-оцінена відповідність змісту документів або текстів інформаційним інтересам споживача.

Пертинентність може оцінити тільки автор запиту, який працює в системі інформаційного

пошуку. Оцінки пертинентности відрізняються від результатів, отриманих на основі оцінок релевантності [9].

### **Висновки**

Інформаційні ресурси Інтернету і наявні в середовищі Інтернет пошукові засоби мають визначену специфіку, яка робить істотний вплив на ефективність пошуку в цьому середовищі. Інтернет – це величезне сховище розподілених оцифрованих даних. Для використання його потенціалу необхідна наявність дієвих інструментів пошуку та обробки даних. Треба відзначити, що єдиної оптимальної схеми пошуку інформації в Інтернеті не існує. Залежно від специфіки потрібної інформації, можна використовувати відповідні пошукові засоби та інструменти. Від того, як грамотно будуть підібрані пошукові засоби та інструменти, залежить якість результатів пошуку.

### **Список літератури**

1. IDC: *Объем информации в интернете удваивается каждые полтора года [Электронный ресурс]. – Режим доступа к ресурсу: <http://www.securitylab.ru/news/379852.php>*
2. *Профессиональный поиск информации в интернете [Электронный ресурс]. – Режим доступа к ресурсу: <http://textbook.vadimstepanov.ru/chapter2/glava2.html>*
3. *Организация ЭВМ и систем [Электронный ресурс]. – Режим доступа к ресурсу: [http://paralichka85.px6.ru/10net/glava10\\_5.htm](http://paralichka85.px6.ru/10net/glava10_5.htm)*
4. Ландэ Д.В., Снарский А.А., Безсуднов И.В., *Интернетика // Ландэ Д.В., Снарский А.А., Безсуднов И.В. –М.: Либроком (Editorial URSS), 2009. - 264 с.*
5. *Информационный поиск в Интернет [Электронный ресурс]. – Режим доступа к ресурсу: <http://synthesis.ipi.ac.ru/sigmod/seminar/s20000330>*
6. *Методи пошуку інформації в мережі інтернет. Інформаційно-пошукові системи [Электронный ресурс]. – Режим доступа к ресурсу: <http://crypto.pp.ua/2011/01/spajdery/>*
7. *Поиск информации в сети Интернет [Электронный ресурс]. – Режим доступа к ресурсу: <http://www.kursach.com!/inforactehnolog/4.6.5.htm>*
8. *Поисковые Интернет журналы [Электронный ресурс]. – Режим доступа к ресурсу: <http://ref.rushkolnik.ru/v45194/>*
9. *Вопросы эффективности поиска информации в интернете и профессиональных базах [Электронный ресурс]. – Режим доступа к ресурсу: <http://mir-masari.narod.ru/0016.html>*

**Рецензент:** д-р техн. наук, проф. И.В. Рубан, Харьковский университет Воздушных Сил им. И. Кожедуба, Харьков.

**Автори:**

**ПАРФЬОНОВ Юрій Едуардович**

Харківський національний економічний університет імені Семена Кузнеця, кандидат технічних наук, старший

науковий співробітник, доцент кафедри інформаційних систем. E-mail – Yuri.Parfonov@m.hneu.edu.ua

**МОРОЗОВ Олександр Ігорович**

Харківський національний економічний університет імені Семена Кузнеця. E-mail – Oleksii.Morozov@m.hneu.edu

#### АНАЛИЗ МЕТОДОВ ИНФОРМАЦИОННОГО ПОИСКА В СЕТИ ИНТЕРНЕТ

Ю.Э. Парфьонов, А.И. Морозов

*В статье рассмотрены основные методы информационного поиска в сети Интернет. Проведен обзор исследований по информационному поиску и определены основные его задачи. Рассмотрены преимущества и недостатки поисковых машин и приведены показатели эффективности поиска информации в Интернете. Сделаны выводы о необходимости использования определенных поисковых средств для получения качественных результатов поиска.*

**Ключевые слова:** *информационный поиск, поисковая машина, Интернет, веб-страница, релевантность.*

#### ANALYSIS OF METHODS OF INFORMATION RETRIEVAL IN THE INTERNET

Y. Parfyonov, A. Morozov

*The article describes the main methods of information retrieval in the Internet. A review of researches on information retrieval is considered and its basic tasks are defined. Advantages and disadvantages of search engines are discussed and performance indicators of information retrieval in the Internet are analyzed. The article concludes that we need to use certain search tools for getting high-quality search results.*

**Keywords:** *information retrieval, search engine, Internet, web page, relevance.*