УДК 004.738.52

О. А. Бойченко
Інститут проблем реєстрації інформації НАН України
вул. М. Шпака, 2, 03113 Київ, Україна
e-mail: boy@cki.ipri.kiev.ua

Створення системи пошуку інформації у корпоративній мережі

Розглянуто особливості пошуку інформації для забезпечення аналітичної діяльності користувачів корпоративних мереж. Запропоновано структуру системи пошуку та методологію її створення, які дозволяють організувати пошук як в межах мережі, так і в Internet.

Ключові слова: *система пошуку, корпоративна комп'ютерна мережа, пошук інформації.*

Сучасні корпоративні комп'ютерні мережі (КМ), переважно побудовані на базі web-технологій, які зарекомендували себе як найбільш прийнятне на сьогодні рішення. Аналітична діяльність вимагає практично миттєвого доступу до багатьох джерел даних, які можуть бути розміщені як у межах корпоративної інформаційно-аналітичної системи, так і в Internet. Створення системи пошуку передбачає вирішення низки завдань, таких як аналіз середовища та формування вимог до системи, моделювання навантаження та продуктивності системи, аналіз ефективності та прогнозованих витрат. Вибір структури системи пошуку повинен залежати від ряду особливостей КМ та забезпечувати виконання основних вимог.

У статті розглядається процес впровадження системи пошуку інформації для вирішення задач пошуку інформації, розташованої як в межах КМ, так і в Internet.

Стрімке зростання обсягів даних, які циркулюють в КМ та Internet спонукає наукові кола та виробників програмного забезпечення до активних досліджень та розробки систем пошуку інформації. В якості основних напрямків досліджень, на думку автора, можна виділити:

— дослідження ефективності роботи та продуктивності систем пошуку інформації [1, 3, 9];

— розробку нових алгоритмів обробки документів та аналізу текстів [6, 7, 12];

— розробку нових архітектур систем пошуку [10, 5, 11].

Одним з основних напрямків у розробці архітектур систем пошуку є розробка розподілених систем, що дозволяє досягти підвищення їх продуктивності та надійності [10, 11].

Моделі інформаційного пошуку поділяються на ряд класів [6]:

1) теоретико-множинні моделі, які базуються на теорії множин;

2) імовірнісні моделі, які базуються на теорії ймовірностей. Наприклад, для оцінки релевантності документа запиту користувача використовується ймовірність того, що користувач визнає документ істинно релевантним;

3) алгебраїчні моделі, які використовуються для опису документів і запитів множини векторів у багатомірному просторі. Каркасом для таких моделей виступають алгебраїчні методи;

4) гібридні моделі, які поєднують властивості вищеназваних класів моделей.

Результати досліджень знаходять втілення у багатьох програмних продуктах, призначених для використання в обмеженому просторі корпоративних систем, а також в Internet.

В таблиці наведені популярні пошукові системи для Inernet.

Глобальні	Російські	Вітчизняні
http://www.google.com	http://www.yandex.ru	http://meta.ua для web-серверів
http://www.alltheweb.com	http://www.rambler.ru	http://uaport.net
http://www.altavista.com	http://www.aport.ru	та http://infostream.com.ua
http://www.yahoo.com		для серверів новин
http://www.msn.com		
http://www.aol.com		
http://www.lycos.com		

Для створення внутрішнього сегмента системи пошуку існує велика кількість програмних продуктів, у тому числі, розроблених найкрупнішими виробниками:

а) Coveo Enterprise Search (www.coveo.com);

б) Oracle text (oracle.com);

в) Sharepoint search (microsoft.com);

г) Google appliance (google.com);

д) Autonomy Knowledge Server (www.autonomy.com).

Слід виділити вітчизняні продукти [8]:

а) MetaSe (meta.com.ua);

б) Dvygun Smart Server (www.dvygun.com);

в) MTSearch.NET (www.aomt.kiev.ua).

Перевагою останніх є початкова підтримка україномовних документів для різних кодувань.

Розглянуті програмні рішення можуть бути успішно використані при організації пошуку в КМ в якості блоків системи пошуку, структурна організація якої та етапи створення будуть розглянуті нижче.

При створенні системи пошуку інформації мають бути вирішені наступні задачі:

— забезпечення пошуку інформації, розташованої на внутрішніх серверах КМ;

— інформування користувачів щодо появи актуальної інформації саме на тих серверах Internet, які є найбільш цікавими для користувача в плані інформаційного наповнення;

— забезпечення роботи з контрольованим набором джерел;

— можливість аналізу інформаційних потреб аналітиків;

— забезпечення захисту інформації.

Пошук на внутрішніх серверах КМ передбачає як безпосередній доступ користувачів до індексів кожного окремого сервера, так і впровадження внутрішньої системи пошуку, яка дозволяє забезпечити централізовану індексацію вмісту серверів даних.

Пошук інформації в Internet передбачає обробку стабільної та динамічної складових Internet. Стабільна складова містить інформацію «довгострокового» плану, наприклад, архіви, колекції, галереї, просто статичні сторінки, які не змінюються. Динамічну складову формують ресурси, які постійно поновлюються.

Для вирішення своїх функціональних задач, система пошуку інформації повинна включати наступні компоненти (рис. 1):

— підсистему індексації КМ;

— БД індексів;

— підсистему внутрішнього пошуку;

— підсистему моніторингу Internet.

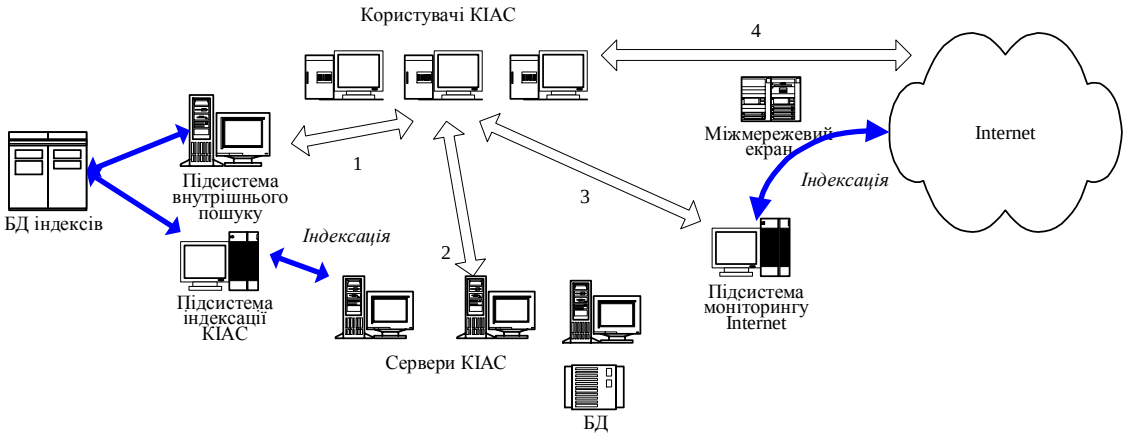

Рис. 1. Організація індексації та пошуку інформації в КМ: 1 — пошук інформації за допомогою підсистеми внутрішнього пошуку: 2 — пошук інформації безпосередньо на серверах КМ; 3 — пошук інформації за допомогою підсистеми моніторингу Internet; 4 — пошук інформації безпосередньо на серверах Internet

Підсистема індексації КМ складається з одного або кількох програмних модулів, кожен з яких індексує певну область рівня даних.

БД індексів забезпечує зберігання індексної інформації.

Підсистема внутрішнього пошуку виконує обробку запитів користувачів КМ та пошук потрібної їм інформації. На рівні даних знаходяться інформаційні сервери, на яких розміщуються бази даних та файлові сховища з файлами різних типів: гіпертекстові, мультимедійні, архівні та ін. На рівні клієнтів знаходяться користу-

вачі, які генерують запити на пошук необхідної їм інформації, використовуючи для цього стандартні засоби перегляду.

Підсистема моніторингу Internet забезпечує постійну індексацію визначеної множини Internet-серверів та надання користувачам найактуальнішої інформації.

Для забезпечення підвищеного рівня захищеності внутрішня мережа КМ може не мати прямого підключення до Internet (рис. 2). У такому випадку для пошуку інформації в Internet необхідно виділити групу робочих місць, з яких користувачі зможуть отримати доступ безпосередньо до серверів Internet та до підсистеми моніторингу.

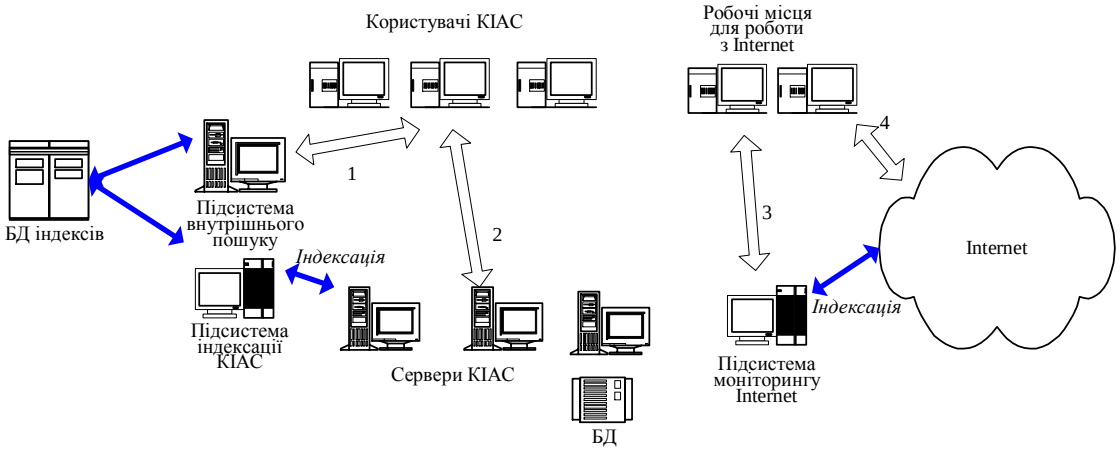

Рис. 2. Організація індексації та пошуку інформації в КМ у випадку розділення внутрішньої та зовнішньої мереж: 1 — пошук інформації за допомогою підсистеми внутрішнього пошуку; 2 — пошук інформації безпосередньо на серверах КМ; 3 — пошук інформації з робочих місць для роботи з Internet за допомогою підсистеми моніторингу Internet; 4 — пошук інформації з робочих місць для роботи з Internet безпосередньо на серверах Internet

Оскільки системи пошуку в КМ слід віднести до класу програмно-апаратних систем, побудованих за web-технологією, то при впровадженні систем пошуку можуть бути використані методи, які застосовуються для планування web-сервісів [1].

При цьому необхідно врахувати наступні особливості системи пошуку:

1) структура системи пошуку включає ряд елементів, частина з яких працює у режимі клієнта, частина — у режимі сервера;

2) елементи системи пошуку є у більшості випадків розподілені в різних хостах мережі, або навіть у різних мережах;

3) використання в роботі досить складних алгоритмів (для обробки документів, для розподілу навантаження між окремими програмними модулями тощо).

Таким чином, основними етапами при проектуванні системи пошуку є наступні:

— аналіз середовища КМ та формування вимог до системи;

— вибір архітектури системи пошуку;

— розклад та прогнозування навантаження;

— моделювання продуктивності;

— аналіз ефективності;

— налаштування та випробовування.

Аналіз середовища та формування вимог передбачає ретельне вивчення програмно-апаратних засобів КМ, інформаційне середовище та регламенти роботи. Особливо важливим для планування системи пошуку є визначення лінгвістичного та інформаційного забезпечення: формати файлів, мова, обсяг тощо. Також необхідно визначитися з вимогами до забезпечення конфіденційності, цілісності та доступності інформації.

Виходячи з вищеназваних задач, сформулюємо основні вимоги до системи пошуку.

1. Підсистема індексації повинна надавати можливість вносити нові документи на індексацію або переіндексацію вже існуючих документів.

2. Можливість контекстного пошуку по документах, які мають найбільш поширені формати документів MS Word (*.doc та *.rtf), MS Excel, txt, html, документи в архівах (*.zip, *.rar).

3. Підсистема відображення результатів пошуку повинна мати можливість сортування результатів пошуку за датою (у прямому та зворотному порядку), за релевантністю, за групами документів, за серверами.

4. Можливість автоматичного оновлення індексу.

5. Підсистема індексації повинна забезпечити режим негайного індексування (поза регламентом) документів, надісланих адміністратором.

6. Автоматичне визначення адрес документів для їх подальшого індексування.

7. Можливість налаштування часового інтервалу між зверненням до документів для уникнення надмірного завантаження підсистеми індексації.

8. Автоматичне розпізнавання мови і типу кодової сторінки документа. У вітчизняних КМ повинні підтримуватися кодові сторінки Windows 1251, KOI-8, Unicode.

9. Розпізнавання форматування документів для врахування при індексації та відображенні.

10. Розпізнавання дублікатів документів.

11. Автоматичне відстеження зміни документів або появи нових (при наявності посилань на них), що гарантує постійну актуальність індексу.

12. Відображення в результатах пошуку наступних параметрів знайдених документів:

— назви документів;

— цитати релевантного фрагмента з виділеними ключовими словами запиту;

— адреси документів і його дублікатів, якщо такі виявлені;

— дати створення документів або останнього поновлення документа;

— кодові сторінки документів;

— розмір документів.

13. Пошук повинен здійснюватися з використанням модулів морфологічного аналізу для української, російської, англійської та інших мов, які використовуються у певній КМ.

14. Мова запитів повинна забезпечувати:

— можливість пошуку точної фрази;

— підтримку логічних операторів: ТА, ЧИ, НІ;

— пошук з усіканням;

— пошук за граматичними формами слів;

— пошук неологізмів, абревіатур, прізвищ тощо;

— пошук за назвою документа.

15. Сумісність із Web-серверами, які підтримують різні технології формування динамічних сторінок.

16. Робота з системою пошуку за допомогою стандартних web-браузерів (Internet Explorer версії не нижче 5.0, Netscafe Navigator версії не нижче 6-х, або інтернет-браузери, в основі яких лежить Mozilla 1-х, тобто Mozilla Suite 1.x.x, Mozilla Phoenix/Firebird/Firefox 0.6 і вище, або Opera версії вище 7.0).

17. Можливість віддаленого адміністрування корпоративної пошукової системи та підсистеми моніторингу.

18. Інтерфейс адміністрування повинен надати адміністратору можливість керувати пошуковим сервісом за наступними критеріями:

— задавати список стартових адрес;

— адмініструвати інтенсивність індексації сервера;

— адмініструвати пошук за такими типами як каталоги і файли (за розширенням);

— накладати заборону на індексування окремих документів або каталогів.

Вибір архітектури системи пошуку повинен здійснюватися з урахуванням вимог системності: забезпечення цільового призначення системи пошуку, сумісності з існуючим програмним та апаратним забезпеченням КМ, модульності та цілісності системи, узгодженості та збалансованості функціональних можливостей системи пошуку з іншими елементами КМ.

Розклад та прогнозування навантаження передбачає виділення із загального робочого навантаження на систему пошуку її окремих складових:

— кількості документів, які мають бути проіндексовані;

— кількості запитів користувачів до підсистеми пошуку;

— середнього обсягу файлу.

Для прогнозування навантаження використовується модель робочого навантаження (Workload Model) [3].

Моделювання продуктивності передбачає прогнозування продуктивності системи для заданих параметрів, серед яких слід виділити параметри системи, що визначаються вибраною архітектурою програмно-апаратних засобів, та параметри робочого навантаження.

Основними показниками продуктивності системи пошуку є наступні [1]:

— коефіцієнт використання серверів;

— коефіцієнт готовності серверів;

— час відгуку серверів.

Систему пошуку можна розглядати як систему масового обслуговування з кінцевою чергою [2], де λ — швидкість надходження запитів до пошукової системи (запитів/с); μ — швидкість обробки запитів до пошукової системи (запитів/с),

$\mu \neq \lambda$; — максимальний розмір черги на обслуговування; k — кількість запитів у черзі.

Для систем з кінцевою чергою вищенаведені показники можна визначити за допомогою наступних формул.

1. Час обробки системою пошуку k запитів:

$$p = \frac{1 - \lambda / \mu}{1 - (\lambda / \mu)^{W+1}} (\lambda / \mu)^k , \tag{1}$$

де, $k = 0,\ldots, W$.

2. Коефіцієнт готовності серверів системи пошуку:

$$U = \frac{(\lambda / \mu)[1 - (\lambda / \mu)^W}{1 - (\lambda / \mu)^{W+1}} (\lambda / \mu)^k . \tag{2}$$

3. Середня продуктивність системи пошуку:

$$X = U \times \mu. \tag{3}$$

4. Середня кількість запитів у системі:

$$\overline{N} = \frac{(\lambda / \mu)[W(\lambda / \mu)^{W+1} - (W+1)(\lambda / \mu)^W + 1]}{(1 - (\lambda / \mu)^{W+1})(1 - \lambda / \mu)} (\lambda / \mu)^k . \tag{4}$$

5. Середній час відгуку серверів системи пошуку:

$$R = \overline{N} / X. \tag{5}$$

Аналіз ефективності використовує результати моделювання продуктивності та моделювання затрат. Модель витрат для системи пошуку повинна врахувати:

— витрати на отримання або розробку програмного забезпечення системи пошуку та додаткові програми (сервери БД, операційні системи, засоби захисту інформації);

— витрати на апаратне забезпечення;

— витрати на телекомунікації (включаючи плату за отримання послуг провайдера Internet);

— витрати на доопрацювання та впровадження.

Налаштування та випробовування системи пошуку дозволяють виявити непередбачені ефекти та недоліки системи.

Запропонована методологія впровадження системи пошуку інформації базується на ретельному аналізі функціональних вимог та планових вимог до системи пошуку та передбачає використання ряду математичних моделей для прогнозування її робочих характеристик.

При реалізації етапів планування є можливим повернення до попереднього етапу у разі виявлення неможливості виконання сформованих раніше вимог.

1. *Менаске Д., Алмейда В.* Производительность Web-служб. Анализ, оценка и планирование. — СПб: ДиаСофтЮП, 2003. — 480 с.

2. *Вентцель Е.С.* Исследование операций. — М.: Советское радио, 1972. — 552 с.

3. *Menasce D., Almeida V., Riedi R., Peligrelliy F., Fonseca R., Wagner M.Jr.* Analyzing Web Robots and Their Impact on Caching. — On line: http://www.cs.bu.edu/techreports/2001-017-wcw01-proceedings/101_almeida.pdf.

4. *Мизин И.А., Богатырев В.А., Кулешов А.П.* Сети коммутации пакетов. — М.: Радио и связь, 1986.

5. *Бойченко О.А.* Про організацію систем пошуку інформації в комп'ютерних мережах // Реєстрація, зберігання і оброб. даних. — 1999. — Т. 1, № 3–4. — С. 45–50.

6. *Некрестьянов И.С.* Тематико-ориентированные методы информационного поиска. Дис... канд. физ.-мат. наук. — On line: http://meta.math.spbu.ru/~igor/thesis/thesis.html.

7. *Ландэ Д.В.* Глубинный анализ текстов. Технология эффективного анализа текстовых данных // CHIP Ukraine. — 2003. — № 10.

8. *Дериев И.* Поисковые системы уровня организации // Компьютерное Обозрение. — 2004. — № 50.

9. *Khoussainov R., Kushmerick N.* Optimizing Performance of Competing Search Engines in Heterogeneous Web Environments // ECML-2003. — Dubrovnik (Croatia). — 2003. — On line: http://www.bridgeport.edu/sed/includes/NEASC%20CSE%20Faculty%20Ativities%20Nov%203%202004.pdf.

10. *Heydon, A. and Najork, M.* Mercator: A Scalable, Extensible Web Crawler // Word Wide Web J. — 1999, Dec. — Vol. 2, N 4. — P. 219–229.

11. *Kasom Koht-arsa*. High Performance Cluster-Based Web Spiders: Master Thesis. — Graduate School. Kasetsart University, 2003. — On line: http://anres.cpe.ku.ac.th/pub/thesis-spider.pdf

12. *Davidov D., Markovitch Sh*. Multiple-Goal Search Algorithms and their Application to Web Crawling. — Haifa (Israel): Computer Science Department Technion.

Надійшла до редакції 05.05.2005