УДК 004.738.5

В.В. ВОЙТКО, П.В. БОДНАР, С.А. ЯРЕМКО
Вінницький національний технічний університет

МЕТОДИ ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ ПОШУКОВИХ ПРОЦЕСІВ У ГЛОБАЛЬНІЙ МЕРЕЖІ

Визначено основні етапи пошукового процесу, розроблено методи онтологічно-орієнтованого та словникового пошуку, які дозволяють підвищити ефективність пошукових процесів у глобальній мережі, проведено аналіз їх особливостей, визначено переваги використання.

Определены основные этапы поискового процесса, разработаны методы онтологически-ориентированного и словарного поиска, которые позволяют повысить эффективность поисковых процессов в глобальной сети, проанализированы их особенности, определены преимущества использования.

Main stages were identified search process, reviewed and analyzed ontologically-oriented and vocabulary search methods, which allow to increase the efficiency of search processes in the global network, given their characteristics and advantages.

Ключові слова: спеціалізований пошук, словниковий пошук, пошукова система, пошуковий запит.

Вступ

Сучасний етап розвитку цивілізації характеризується переходом людства від індустріального суспільства до інформаційного. Інформаційні системи стають основними засобами вирішення задач різних видів діяльності і формують галузь індустрії інформаційних технологій, що досить бурхливо розвивається сьогодні. Одним з найбільш яскравих явищ процесу інформатизації є виникнення і розвиток глобальної комп'ютерної мережі. Порівняно з традиційними методами глобальна база даних значно розширює можливості отримання необхідної інформації.

У процесі розвитку всесвітньої глобальної мережі Інтернет проблема пошуку потрібних інформаційних ресурсів набуває актуальності і потребує розробки та удосконалення пошукових систем. Сьогодні проблема пошуку потрібної інформації у глобальній мережі залишається не до кінця вирішеною. Сучасні пошукові системи постійно стикаються з низкою технологічних проблем, пов’язаних з динамічним ростом загальних обсягів інформаційних ресурсів, переходом до нових форматів (XML), збільшенням кількості нетекстових матеріалів тощо [1]. Як наслідок означених недоліків, результати виконання запиту не завжди відповідають шуканим даним. Це призводить до подальшого переформулювання користувацьких запитів, зміни пошукових систем, що в свою чергу потребує додаткових часових затрат. Тому в умовах прискорених темпів інформатизації все більшої актуальності набуває проблема впровадження спеціалізованих пошукових систем, орієнтованих на вузькопрофільний пошук інформаційних ресурсів.

Постановка завдання

Метою роботи є підвищення ефективності пошукових процесів у глобальній мережі.

Об'єктом дослідження постають пошукові системи.

Під предметом дослідження розуміємо методи та засоби реалізації мережевих пошукових процесів.

Головними задачами вбачаємо розробку та впровадження методів пошуку, які дозволять підвищити швидкість, релевантність пошукових процесів, зменшити ресурсні затрати на їх виконання.

1. Визначення етапів пошукового процесу

Зі збільшенням об’єму інформаційних ресурсів постійно зростають вимоги до сучасних пошукових систем. Серед головних з них виділяють швидкість пошуку, релевантність, ресурсні затрати, якість отриманих даних та гнучкість пошукових процесів.

У загальному випадку пошукові системи складаються з п'яти окремих взаємопов'язаних програмних компонентів:

1. Пошуковий павук – подібний до програми браузера, що завантажує Web-сторінки. Павук сам зв'язується з інформаційною базою даних і переглядає Web-сторінки глобальної мережі.

2. Мандрівний павук – програма розпізнає сторінку і відшукує на ній усі посилання.

3. Індексатор – програма, що розділяє сторінки на частини і проводить аналіз кожної з них. Індексатор аналізує заголовки сторінок, посилання, текст та інші структурні елементи.

4. Бази даних зберігають інформацію після аналізу та подальшої обробки пошуковою системою.

5. Система видачі результатів обирає відповіді, які задовольняють запит користувача [2].

Схему процесу додавання нових даних до реєстру пошукової системи зображено на рисунку 1.

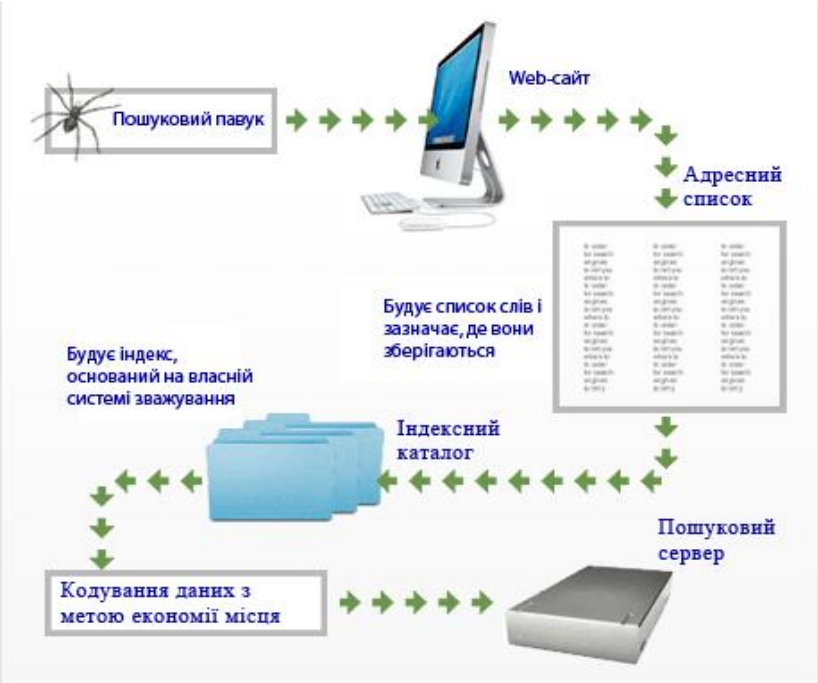

Рисунок 1 – Схема процесу додавання даних до пошукової системи

Пошуковий процес, у свою чергу, складається з десяти базових етапів:

- фіксації інформаційної потреби природною мовою користувача;
- вибору потрібних пошукових сервісів мережі і точної формалізація запису інформаційної потреби символьним апаратом обраної інформаційно-пошукової мови;
- виконання створених запитів;
- вибірки і попередньої обробки отриманих списків посилань на документи;
- звернення за обраними адресами до шуканих документів;
- попередній перегляд вмісту знайдених документів;
- збереження релевантних документів для подальшого вивчення;
- витяг з релевантних документів посилань для розширення запиту;
- вивчення всього масиву збережених документів;

– якщо інформаційна потреба не повністю задоволена, то відбувається повернення до першого етапу реалізації пошукового процесу.

2. Метод онтологічно-орієнтованого пошуку

Одним з методів підвищення швидкості виконання пошукових процесів є використання спеціалізованих онтологічно-орієнтованих пошукових систем. Такі системи характеризуються пошуком потрібної інформації в базах даних обраної предметної області та відсутністю розрізнених різнотипних даних. Усі матеріали певного спеціалізованого напрямку мають однотипну структуру, за якою їх легко розділити на різні класифікаційні групи, що дозволяє зменшити формат пошуку та скоротити часові затрати пошукових операцій. Таким чином, спеціалізовані пошукові системи дозволяють здійснити пошук за певними чіткими критеріями у конкретно обраних категоріях, зменшити надлишковість пошукових операцій, оптимізувати ресурсні затрати і час пошуку потрібної інформації шляхом ідентифікованого обмеження бази пошукових процесів [3]. Тому, пошуковий запит із заданими характеристиками, який відповідає інформаційним ресурсам *i*-тої гілки каталогу із загальною кількістю гілок першого рівня *n* та однаковою кількістю даних, акумульованих кожною гілкою, виконається в *n-1* разів швидше, ніж аналогічний запит у звичайній пошуковій системі з тією ж кількістю даних. Такий підхід орієнтований на використання пошукових фільтрів, які реалізують пошук в обраних гілках дерева даних. Це дозволяє значно збільшити швидкість пошуку потрібної інформації, підвищити релевантність та зменшити ресурсні затрати на виконання пошукового процесу [4].

У спеціалізованій пошуковій системі значно зменшується кількість надлишкових запитів, оскільки користувач заздалегідь орієнтується на отримання результатів з конкретної інформаційної групи, що відкидає потребу уточнення запиту за типом шуканої інформації у базі різногалузевих даних. У свою чергу такий пошук дозволяє зменшення навантаження на сервер пошукової системи, оскільки він розцінює кожне уточнення як новий запит і виконує його з тими ж ресурсними затратами, як і початковий [5].

3. Метод словникового пошуку

Ще одним методом швидкого пошуку у великих об’ємах текстових даних вбачаємо словниковий метод, який нагадує пошук слова у словнику. Він дозволяє отримати якість результатів пошукових операцій, наближених до якості методу повного перебору, та забезпечує швидкість пошуку з часовими затримками, необхідними для проходження гілки дерева графа. В словниковому методі час пошуку практично не залежить від об’єму матеріалів, за якими виконується пошуковий запит. Відомо, що в мовних засобах для різних стилів та жанрів існує близько 40000 найчастіше вживаних слів, а пересічна людина має словниковий запас близько 5000-10000 слів [6].

Метод словникового пошуку орієнтований на роботу з обмеженою кількістю записів у словнику, незалежно від загальної кількості матеріалів. Метод базується на групуванні значущих слів робочих матеріалів (інформаційних баз даних електронних книг, журналів, web-сторінок) у так званий «словник». Слова додаються виключно в алфавітному порядку без їх повторень з урахуванням положення кожного наступного символу слова в алфавіті. При заповненні словника для кожного символу формується масив службової інформації про наступні літери слова. Масив акумулює список літер (для українського алфавіту – від літери А до Я) з відповідними адресами меж розміщення кожної літери в словнику (рис. 2).

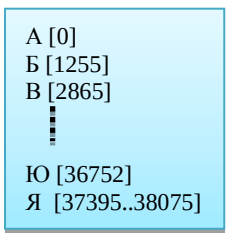

Рисунок 2 – Приклад запису службової інформації алфавіту словника

Цифровий масив, поданий у квадратних дужках, вказує на початкову адресу слів, у яких обрана літера є наступною шуканою. Адреси слів останньої літери списку записуються через знак "..", що позначає проміжок, на якому розміщено слова, котрі мають вказану літеру. Таким чином, адресний проміжок літери X при записах словникового алфавіту "X [N]", розміщеного перед записом "Y [M]", знаходиться за виразом (1):

$$A_X = [N..M-1], \tag{1}$$

де A_X – адресний проміжок розміщення слів з літерою X;

N – початкова адреса слів з літерою X;

M – початкова адреса слів з літерою Y.

Кожне слово словника містить посилання на джерело, з якого воно було додане, а також конкретне місце розташування в цьому джерелі (рис. 3). Тобто, якщо слово існує в словнику, то його можна відшукати в оригіналі, перейшовши за отриманими адресами.

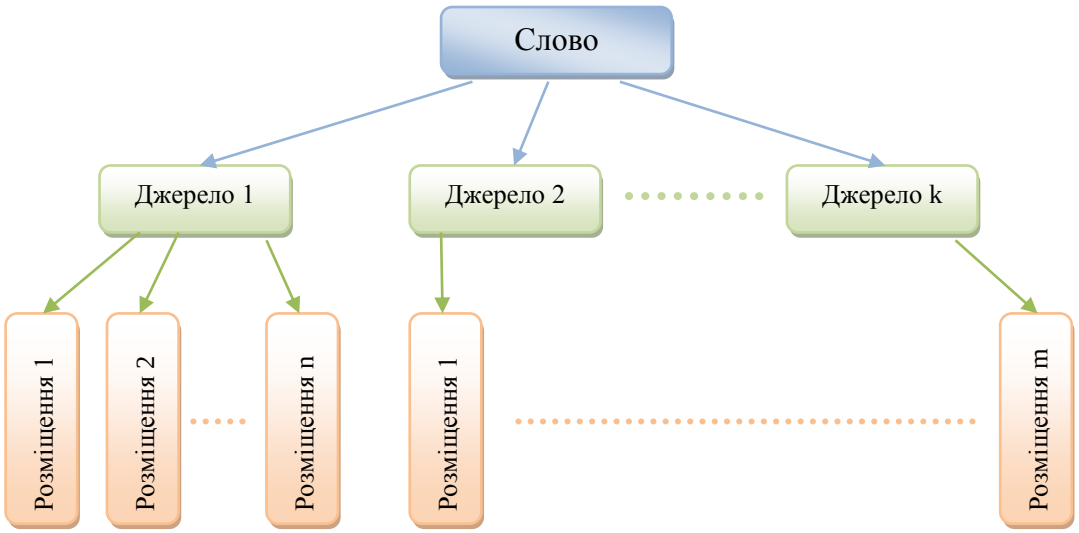

Рисунок 3 – Загальна структура адресації слова

Запропонована структура словника орієнтована на організацію зв'язок слів за їх класифікацією, як, наприклад, за визначенням синонімів, антонімів та інших варіантів словосполучень зі змістовою складовою. Такий підхід дозволить розширити структуру пошукового запиту, наблизивши його до природної мови людини [7].

Додавання слів з нового тексту здійснюється шляхом аналізу кожного слова та внесення до словника адреси джерела і місця розміщення в ньому значущого слова. Нові слова вводять у словник за алфавітним порядком. Словниковий метод передбачає швидкий пошук потрібного слова та забезпечує можливість отримання додаткової інформації за запитом користувача. Таким чином, словник акумулює службову інформацію, корисну для реалізації швидкодіючих пошукових процесів, наприклад, щільність шуканих слів (кількість шуканих слів, що містяться на одній сторінці тексту), перелік сторінок з високою щільністю шуканих слів і т.п.

Граф-схема узагальненого алгоритму додавання матеріалу до словника зображена на рисунку 4.

Змінна N показує кількість слів у тексті, $W[i]$ – i-те слово матеріалу, D – словниковий масив даних, $f(i)$ – довжина i-го слова, A – поточний адресний простір, $A[j]$ – адресний простір j-го символу слова підмножини A. При додаванні адреси слова записується ідентифікаційний номер матеріалу та позиція слова в ньому (у вигляді номера стрічки/номера символу).

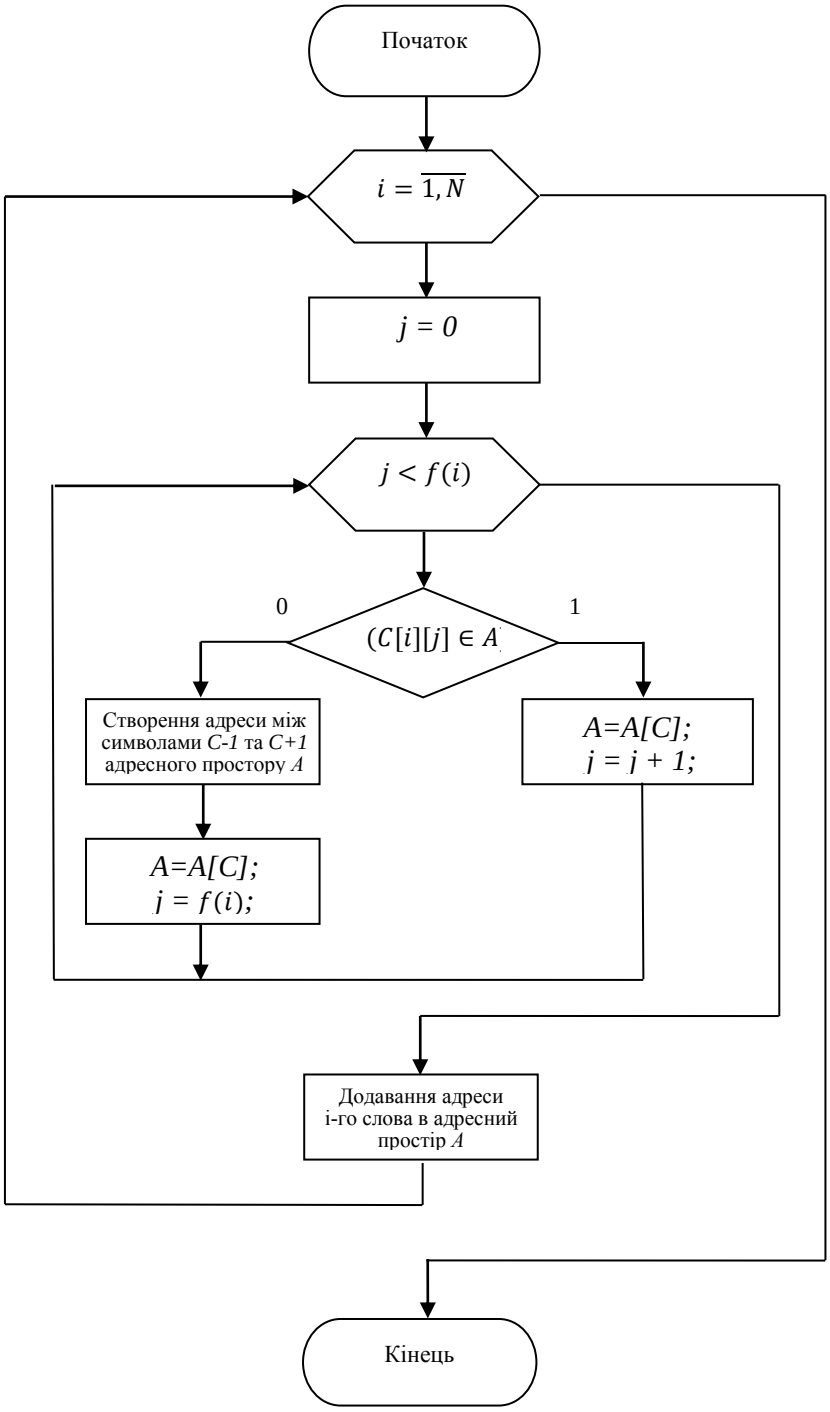

Рисунок 4 – Граф-схема алгоритму додавання матеріалів до словника

Розмір словника, головним чином, залежить від кількості матеріалів та максимального числа слів у них і приблизно рівний загальному розміру усіх матеріалів, доданих до нього. Так у словнику з максимальною кількістю матеріалів n та максимальною кількістю слів у кожному матеріалі m, індекси розміщення слова будуть мати розміри, еквівалентні A' символам (2).

$$A = (\frac{\log_2 n + \log_2 m}{8});$$
$$A' = f(A), \tag{2}$$

де $f(x)$ – оператор округлення до більшого цілого.

Наприклад, для словника з максимальною кількістю матеріалів 16 777 216 та максимальною кількістю слів 4 294 967 296 індекс буде мати такий розмір, як і слово з 7 літер (при кодуванні символів найпоширенішим є восьмибітний код).

Програмна реалізація словникового методу передбачає організацію швидкісних пошукових процесів шляхом символьної ідентифікації слів. Рисунок 5 ілюструє приклад пошуку слова “WORD”. Під a, b, c розуміються числові значення адрес, X – адреса шуканого слова.

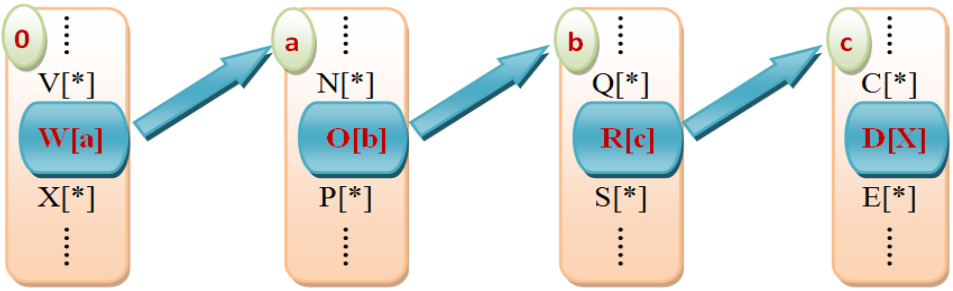

Рисунок 5 – Приклад пошуку за словниковим методом

За знайденою адресою X зчитуються ідентифікаційні номера джерел, за якими складається система посилань, та фіксуються позиції розміщення в тексті всіх шуканих слів. Результати пошукового процесу формують відповідь на пошуковий запит користувача.

Висновки

Таким чином, використання онтологічно-орієнтованого та словникового методів для реалізації пошукових систем сумісно з використанням додаткових засобів покращення пошукового процесу, таких як семантичний та морфологічний розбір, аналіз за структурою словотворення, створює ефективну спеціалізовану пошукову систему, що дозволяє більше, ніж на порядок підвищити швидкість пошуку даних і відповідно зменшити кількість обчислювальних ресурсів на здійснення пошукових операцій. Як наслідок, з’являється можливість зменшення економічних та енергетичних затрат на використання та обслуговування надпотужних дата-центрів шляхом їх заміни спеціалізованими базами даних, орієнтованими на реалізацію онтологічних пошукових процесів.

Література

1. Антонов А.В., Мешков В.С. Современные проблемы поисковых систем и некоторые пути их преодоления. [Электронный ресурс]. Режим доступа: http://www.galaktika-zoom.ru/publications/p01.
2. Гусев В.С. Google. Эффективный поиск. Как работают поисковые машины. – М.: Издательский дом Вильямс, 2006, – С. 39-41.
3. Бевз С., Бурбело С., Боднар П.. Спеціалізовані пошукові системи для глобальної мережі. / Матеріали V Міжнародної науково-технічної конференції «Сучасні проблеми радіоелектроніки, телекомунікацій та приладобудування (СПРТП-2011)». Вінниця, – 19-21 травня 2011, – С. 32-33.
4. Войтко В.В., Бевз С.В., Бурбело С.М., Боднар П.В. Спеціалізовані пошукові системи для глобальної мережі. / Електронні ресурси та технології: створення, використання, доступ. Збірник матеріалів Міжнародної науково-практичної Інтернет-конференції. Вінниця, – 10-17 травня 2011, – С. 60-62.
5. Гультяев А. Самое главное о... Поиск в Интернете. 2-е издание. /Тематический поиск. М:. Питер, 2006, – С. 56-70.
6. Карпухин С. Сколько слов в русском языке? – Наука и жизнь. №11, 2004.
7. Романенко В. Н.,. Никитина Г. В. Сетевой информационный поиск. /Поиск при запросах на естественном языке. – М.: «Профессия», 2005, – С. 105-112.