УДК 004.421
Яцишин В.П. – ст. гр. СНм-51
Тернопільський національний технічний університет імені Івана Пулюя

АЛГОРИТМИ ПОШУКОВИХ СИСТЕМ

Науковий керівник: асистент Прошин С.Ю.

Сучасні пошукові системи розрізняються реалізованими в них алгоритмами і структурами даних. Можна виділити 4 основні алгоритми, на яких засновані всі сучасні пошукові системи:

- Прямий пошук.
- Інвертований список.
- Суфіксні дерева.
- Сигнатури.

Перший, найпростіший алгоритм, не припускає попереднього етапу індексації документа, і пошук ведеться шляхом послідовного перегляду документів. Очевидно, прямий перегляд великих об'ємів тексту – досить повільний прийом пошуку, але, проте, він іноді використовується навіть в пошукових системах для Інтернету, наприклад, норвезька система www.fastsearch.com використовує саме прямий метод. Для прямого пошуку розроблені не тільки найпростіші алгоритми по методу «грубої сили», але і більш ефективні, включаючі можливість пошуку за шаблоном, такі, як різні модифікації алгоритму Бойера-Мура. Крім того, більшість програм комбінує індексний пошук для знаходження блоку тексту з подальшим прямим пошуком усередині цього блоку.

Інвертований список – це прийом зберігання інформації в БД, при якому індексується кожне слово документа, і при цьому зберігається значення його позиції в документі. Наприклад, так: «книга 23, розділ 8, параграф 1, абзац 12, слово 114». У такому разі пошуковий алгоритм зведеться до знаходження слова в БД і видачі посилань на конкретні позиції в документі. Часто застосовують також які-небудь способи упаковки для поля, що береже позицію слова, в найпростішому випадку це може бути зберігання не абсолютної позиції, а відстані від попередньої. Використовуються також алгоритми стиснення типу Хоффмана або LZW, але рідше, оскільки вони трохи збільшують швидкість доступу до даних, та зате сильно навантажують процесор.

Суфіксні дерева – запатентований алгоритм пошукової системи OpenText. Суфіксні дерева, суфіксні масиви (suffix trees, suffix arrays) є індексом, заснованим на занесенні всіх значущих суфіксів тексту в структуру даних, відому як «дерево» (trie). Подібна організація даних використовується, зокрема, відомою україномовною пошуковою системою «Яндекс». Суфіксом в цьому індексі називають будь-який підрядок, що починається з деякої позиції тексту (текст розглядається як один безперервний рядок) і триваючу до його кінця. В реальних програмах довжина суфіксів обмежена, а індексуються тільки значущі позиції – наприклад, початки слів. Цей індекс дозволяє виконувати складніші запити, ніж індекс, побудований на інвертованих файлах. Стосовно задачі повнотекстового пошуку алгоритм суфіксних дерев не є достатньо ефективним. Останнім часом намітилася також тенденція використовування суфіксних дерев в задачах класифікації результатів пошукових запитів і побудови тематичних фільтрів.

Метод сигнатур є перетворенням документа до поблочних таблиць хеш-значень його слів – «сигнатур» і послідовному перегляду сигнатур під час пошуку.