ВІСНИК ЛЬВІВ. УН-ТУ
Серія прикл. матем. інформ.
2011. Вип. 17. С. 22–28

VISNYK LVIV UNIV.
Ser. Appl. Math. Inform.
2011. Is. 17. P. 22–28

УДК 004.632.3

ПОРІВНЯЛЬНИЙ АНАЛІЗ ОПТИМАЛЬНИХ СХЕМ ПОШУКУ ІНФОРМАЦІЇ У ПОСЛІДОВНИХ ФАЙЛАХ БАЗ ДАНИХ ІЗ ВИКОРИСТАННЯМ МЕТОДУ БЛОКОВОГО ПОШУКУ

А. Мельничин, Г. Цегелик

Львівський національний університет імені Івана Франка,
вул. Університетська, 1, Львів, 79000, e-mail: kafmmsep@franko.lviv.ua

Для різних законів розподілу ймовірностей звертання до записів проведено порівняльний аналіз оптимальних схем трьох варіантів пошуку інформації у послідовних файлах баз даних із використанням методу блокового пошуку. За критерій оптимальності прийнято математичне сподівання загального часу, потрібного для пошуку запису у файлі.

Ключові слова: послідовні файли баз даних, блоковий пошук.

1. ВСТУП

У [1–3] досліджено ефективність блокового пошуку записів у послідовних файлах баз даних, які містяться в зовнішній пам'яті однопроцесорних ЕОМ, залежно від вибраного варіанта пошуку записів і розподілу ймовірностей звертання до записів. Серед варіантів пошуку розглянуто такі:

- послідовне читання блоків записів в основну пам'ять і їхній перегляд за допомогою методу блокового пошуку;
- використання методу блокового пошуку в блоці записів, який попередньо локалізований шляхом читання блоків записів в основну пам'ять і перегляду їхніх останніх записів;
- використання методу блокового пошуку в блоці записів, який попередньо локалізований шляхом читання і перегляду останніх записів кожного блока.

Серед законів розподілу ймовірностей звертання до записів розглянуто рівномірний і такі близькі до реальності нерівномірні розподіли: "бінарний", Зіпфа та узагальнений, частковим випадком якого є розподіл, що наближено задовольняє правило "80-20" [4]. У праці для тих самих законів розподілу ймовірностей звертання до записів проведено порівняльний аналіз оптимальних схем згаданих варіантів пошуку.

2. ФОРМУЛЮВАННЯ ЗАДАЧІ

Розглянемо послідовний файл, який містить N записів. Припустимо, що файл зберігається в зовнішній пам'яті ЕОМ і всі його записи умовно розбиті на n блоків по ms записів у кожному. Якщо використовують блоковий пошук у блоці записів, то його розбивають на s підблоків по m записів у кожному. Введемо позначення:

$a = b + dms$ – час читання блока записів в основну пам'ять, де b, d – деякі сталі; t – час перегляду запису в основній пам'яті; p_i – ймовірність звертання до i-го запису файла; E – математичне сподівання загального часу, потрібного для пошуку запису у файлі.

Проведемо порівняльний аналіз оптимальних схем розглянутих варіантів пошуку записів у файлі для різних законів розподілу ймовірностей звертання до записів.

3. ПЕРШИЙ ВАРІАНТ ПОШУКУ

Припустимо, що пошук записів у файлі відбувається шляхом послідовного читання блоків записів в основну пам'ять і їхнього перегляду за допомогою методу блокового пошуку.

Якщо подати математичне сподівання загального часу, потрібного для пошуку запису у файлі, у вигляді суми математичного сподівання часу, потрібного для пошуку блока, математичного сподівання часу, потрібного для пошуку підблока, і математичного сподівання часу, потрібного для пошуку запису в підблоці, то [3]

$$E = \sum_{i=1}^{n}\left(ai + (i-1)st\right)\sum_{k=1}^{s}\sum_{j=1}^{m} p_{(i-1)ms+(k-1)m+j} + \\ + \sum_{i=1}^{n}\sum_{k=1}^{s} kt \sum_{j=1}^{m} p_{(i-1)ms+(k-1)m+j} + \sum_{i=1}^{n}\sum_{k=1}^{s}\sum_{j=1}^{m} jtp_{(i-1)ms+(k-1)m+j},$$

або

$$E = \sum_{i=1}^{n}\sum_{k=1}^{s}\sum_{j=1}^{m}\left(ai + \left((i-1)s + k + j\right)t\right)p_{(i-1)ms+(k-1)m+j} \quad .$$

Математичне сподівання загального часу, потрібного для пошуку запису у файлі, для різних законів розподілу ймовірностей звертання до записів обчислюється за формулами:

- у випадку рівномірного розподілу ймовірностей

$$E = \frac{1}{2}\left(\left(\frac{N}{sm}+1\right)\left(b+dsm\right)+\left(\frac{N}{m}+m+2\right)t\right);$$

- для "бінарного" розподілу ймовірностей з точністю до нескінченно малої величини

$$E = \frac{2^{sm}}{2^{sm}-1}\left(b+dsm\right)+\left(3-\frac{m-1}{2^{m}-1}\right)t;$$

- для закону Зіпфа з достатньо високою точністю

$$E = \frac{1}{H_N}\left(\left(H_N + \frac{N}{sm} - \frac{1}{2}\ln\frac{N}{sm} - C_1\right)\left(b+dsm\right)+\left(H_N+\frac{N}{m}+(m-1)\left(\frac{1}{2}\ln\frac{N}{m}+C_1\right)\right)t\right),$$

де $C_1 = \frac{1}{2}\ln 2\pi$, $H_N = \sum_{k=1}^{N}\frac{1}{k}$ – частинна сума гармонічного ряду;

- у випадку узагальненого закону розподілу ймовірностей з достатньо високою точністю

$$E = \frac{1}{H_N^{(c)}}\left(\left(H_N^{(c)} + \frac{N^{1-c}}{1-c}\left(\frac{1-c}{2-c}n - \frac{\alpha^{(c)}(n)}{n^{1-c}}\right)\right)\left(b+\frac{dN}{n}\right)+\right.$$

$$+\left(H_N^{(c)}+H_N^{(c-1)}+\frac{N^{2-c}}{1-c}\left(1-\frac{1}{m}\right)+(m-1)\frac{\alpha^{(c)}\left(N/m\right)}{\left(N/m\right)^{1-c}}\right)t\Bigg),$$

де c – будь-який параметр ($0 < c < 1$); $H_N^{(c)}=\sum_{k=1}^{N}\frac{1}{k^c}$ – частинна сума узагальненого гармонічного ряду, $\alpha^{(c)}\left(n\right)=H_n^{(c-1)}-\frac{1}{2-c}n^{2-c}$ – повільно зростаюча функція.

4. ДРУГИЙ ВАРІАНТ ПОШУКУ

Процес пошуку потрібного запису у цьому випадку відбувається так: спочатку локалізується блок, який містить шуканий запис, шляхом послідовного читання блоків записів в основну пам’ять і перегляду їхніх останніх записів. Після цього пошук потрібного запису продовжується в локалізованому блоці за допомогою методу блокового пошуку.

Якщо зобразити математичне сподівання загального часу, потрібного для пошуку запису у файлі, у вигляді суми математичного сподівання часу, потрібного для локалізації блока записів, математичного сподівання часу, потрібного для локалізації підблоку записів, і математичного сподівання часу, потрібного для пошуку запису в локалізованому підблоці, то E виразиться формулою [1]

$$E=\sum_{i=1}^{n}(a+t)i\sum_{k=1}^{s}\sum_{j=1}^{m}p_{(i-1)ms+(k-1)m+j}+\sum_{i=1}^{n}\sum_{k=1}^{s}kt\sum_{j=1}^{m}p_{(i-1)ms+(k-1)m+j}+\sum_{i=1}^{n}\sum_{k=1}^{s}\sum_{j=1}^{m}jtp_{(i-1)ms+(k-1)m+j},$$

або

$$E=\sum_{i=1}^{n}\sum_{k=1}^{s}\sum_{j=1}^{m}\left(ai+\left(i+k+j\right)t\right)p_{(i-1)ms+(k-1)m+j}\ .$$

Математичне сподівання загального часу, потрібного для пошуку запису у файлі, для законів розподілу ймовірностей звертання до записів обчислюється за формулами:

- у випадку рівномірного розподілу ймовірностей

$$E=\frac{1}{2}\left(\left(\frac{N}{ms}+1\right)(b+dms)+\left(\frac{N}{ms}+s+m+3\right)t\right);$$

- для “бінарного” розподілу ймовірностей з точністю до нескінченно малої величини

$$E=\frac{2^{ms}}{2^{ms}-1}(b+dms)+\left(\frac{2^{ms}-s}{2^{ms}-1}+\frac{3\cdot 2^m-m-2}{2^m-1}\right)t;$$

- для закону Зіпфа з достатньо високою точністю

$$E=\frac{1}{H_N}\left(\left(H_N+\frac{N}{ms}-\frac{1}{2}\ln\frac{N}{ms}-C_1\right)(b+dms)+\right.$$
$$\left.+\left(2H_N+\frac{N}{ms}+\frac{1}{2}\left((s+m-2)\ln\frac{N}{ms}+(m-1)\ln s\right)+(s+m-2)C_1\right)t\right);$$

- у випадку узагальненого закону розподілу ймовірностей з достатньо високою точністю

$$E=\frac{1}{H_N^{(c)}}\left(\left(H_N^{(c)}-\frac{N^{1-c}}{1-c}\left(\frac{c-1}{2-c}n+\frac{\alpha^{(c)}(n)}{n^{1-c}}\right)\right)\left(b+d\frac{N}{n}\right)+\right.$$

$$\left.+\left(2H_N^{(c)}+H_N^{(c-1)}+\frac{N^{1-c}}{1-c}\left((N-n)\frac{c-1}{2-c}+\frac{(s-1)\alpha^{(c)}(n)}{n^{1-c}}+\left(\frac{N}{ns}-1\right)\frac{\alpha^{(c)}(ns)}{(ns)^{1-c}}\right)\right)t\right).$$

5. ТРЕТІЙ ВАРІАНТ ПОШУКУ

Пошук потрібного запису відбувається так. Спочатку локалізується блок, який містить шуканий запис, шляхом послідовного читання і перегляду останніх записів кожного блока. Після цього пошук потрібного запису продовжується в локалізованому блоці за допомогою методу блокового пошуку.

Зобразимо математичне сподівання загального часу, потрібного для пошуку запису у файлі, у вигляді суми математичного сподівання часу, потрібного для локалізації блока записів, математичного сподівання часу, потрібного для локалізації підблока записів, і математичного сподівання часу, потрібного для пошуку запису в локалізованому підблоці. Тоді математичне сподівання загального часу, потрібного для пошуку запису у файлі, виразиться формулою [2]

$$E=a+\sum_{i=1}^{n}it_1\sum_{k=1}^{s}\sum_{j=1}^{m}p_{(i-1)ms+(k-1)m+j}+\sum_{i=1}^{n}\sum_{k=1}^{s}kt\sum_{j=1}^{m}p_{(i-1)ms+(k-1)m+j}+\sum_{i=1}^{n}\sum_{k=1}^{s}\sum_{j=1}^{m}jtp_{(i-1)ms+(k-1)m+j},$$

або

$$E=\sum_{i=1}^{n}\sum_{k=1}^{s}\sum_{j=1}^{m}(a+it_1+(k+j)t)p_{(i-1)ms+(k-1)m+j},$$

де $t_1=b+d+t$ – час читання запису в основну пам'ять і його перегляд.

Математичне сподівання загального часу, потрібного для пошуку запису у файлі, для законів розподілу ймовірностей звертання до записів, обчислюють за формулами:

- у випадку рівномірного розподілу ймовірностей

$$E=b+dms+\frac{1}{2}\left(\left(\frac{N}{ms}+1\right)(b+d)+\left(\frac{N}{ms}+s+m+3\right)t\right);$$

- для "бінарного" розподілу ймовірностей з точністю до нескінченно малої величини

$$E=b+dms+\frac{2^{ms}}{2^{ms}-1}(b+d)+\left(4+\frac{s-1}{2^{ms}-1}-\frac{m-1}{2^{m}-1}\right)t;$$

- для закону Зіпфа з достатньо високою точністю

$$E=b+dms+\frac{1}{H_N}\left(\left(H_N+\frac{N}{ms}-\frac{1}{2}\ln\frac{N}{ms}-C_1\right)(b+d)+\right.$$

$$\left.+\left(2H_N+\frac{N}{ms}+\frac{1}{2}\left((s+m-2)\left(\ln\frac{N}{ms}+2C_1\right)+(m-1)\ln s\right)\right)t\right);$$

- у випадку узагальненого закону розподілу ймовірностей з достатньо високою точністю

$$E = b + d\frac{N}{n} + \frac{1}{H_N^{(c)}}\left(\left(H_N^{(c)} - \frac{N^{1-c}}{1-c}\left(\frac{c-1}{2-c}n + \frac{\alpha^{(c)}(n)}{n^{1-c}}\right)\right)(b+d) + \right.$$

$$\left. + \left(2H_N^{(c)} + H_N^{(c-1)} + \frac{N^{1-c}}{1-c}\left((N-n)\frac{c-1}{2-c} + \frac{(s-1)\alpha^{(c)}(n)}{n^{1-c}} + \left(\frac{N}{ns} - 1\right)\frac{\alpha^{(c)}(ns)}{(ns)^{1-c}}\right)\right)t\right).$$

6. ПОРІВНЯЛЬНИЙ АНАЛІЗ

У таблиці для розглянутих варіантів пошуку (з точністю до 0.1) подано оптимальні значення величини E/d для різних законів розподілу ймовірностей звертання до записів для деяких b/d, $t/d = 0.1$ і $N = 10^6$.

Оптимальні значення величини *E/d* для різних законів розподілу ймовірностей звертання до записів для деяких *b/d*, *t/d*=0.1 і *N*=10^6

Варі-анти	*b/d*	Закони розподілу						
		c=0	***c=0.2***	***c=0.4***	***c=0.6***	***c=0.8***	***c=1***	***"Бінарний"***
I	**10**	503267,4	447526,2	377895,2	289147,1	178666,9	70907,3	15,1
	100	510150,1	454007,6	383906,9	294536,7	183199,6	74102,3	107,9
	1000	532222,9	475026,1	403565,2	312385,5	198371,7	84949,1	1011,2
II	**10**	503188,8	447458,2	377840,2	289070,1	178586,9	70886,5	15,2
	100	510065,2	453962,8	383871,7	294486,6	183134,9	74078,1	108,0
	1000	532142,3	474945,1	403489,6	312331,3	198341,4	84925,5	1011,3
III	**10**	4732,2	4462,6	4101,0	3587,2	2819,6	1776,3	25,9
	100	14378,8	13565,5	12474,9	10926,5	8615,6	5478,0	209,0
	1000	46261,6	43706,3	40284,1	35436,9	28225,9	18456,3	2012,3

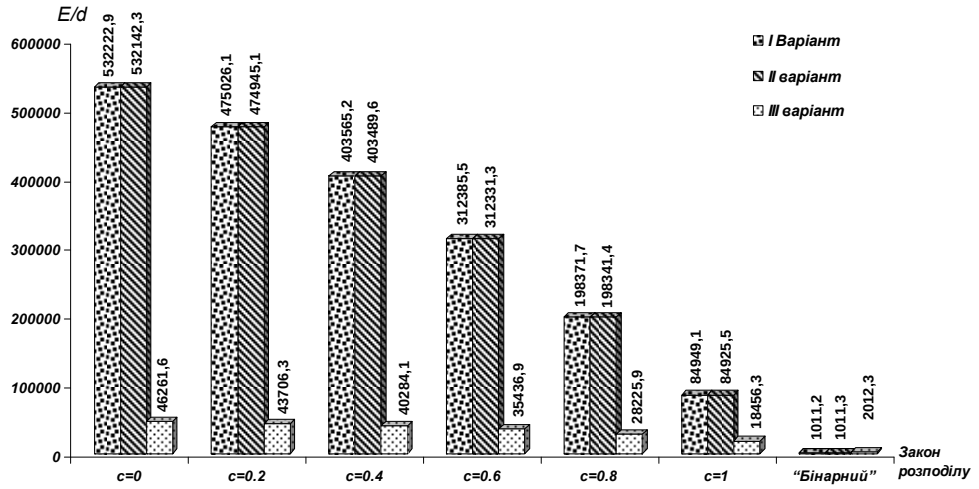
Зауважимо, що у таблиці значенню ***c = 0*** відповідає рівномірний розподіл, а значенню ***c = 1*** – закон Зіпфа.

На рис. показана залежність оптимального значення величини E/d від зміни закону розподілу ймовірностей звертання до записів для розглянутих варіантів пошуку, різних законів розподілу ймовірностей звертання до записів, $b/d = 1000$, $t/d = 0.1$ і $N = 10^6$.

Як видно з рис., зі зміною закону розподілу ймовірностей звертання до записів від рівномірного до "бінарного" оптимальне значення величини *E*/*d* суттєво зменшується. Для всіх законів розподілу, крім "бінарного", третій варіант пошуку значно ефективніший за перший і другий. У випадку "бінарного" закону розподілу перший та другий варіанти майже удвічі ефективніші за третій. Якщо порівнювати за ефективністю перший та другий варіанти, то вони практично однакові.

7. ВИСНОВКИ

Розглянуто три різні варіанти пошуку записів у послідовних файлах із використанням методу блокового пошуку. Для різних законів розподілу ймовірностей звертання до записів (рівномірного, "бінарного", Зіпфа, узагальненого, частковим випадком якого є розподіл, що наближено задовольняє правило "80-20") проведено порівняльний аналіз оптимальних схем цих варіантів. За критерій оптимальності прийнято математичне сподівання загального часу, потрібного для пошуку запису в файлі.



Оптимальні значення величини E/d для різних варіантів пошуку, різних законів розподілу ймовірностей звертання до записів, b/d=1000, t/d=0.1 і N=10^6

ЛІТЕРАТУРА

1. *Мельничин А. В.* Оптимальні стратегії пошуку записів в послідовних файлах баз даних при використанні методу блочного пошуку в локалізованому блоці записів / *А. В. Мельничин, Г. Г. Цегелик* // Наук. вісник Ужгород. ун-ту. Сер. матем. і інформ. – 2009. – Вип. 18. – С. 92–98.
2. *Мельничин А. В.* До побудови оптимальних стратегій пошуку записів у послідовних файлах баз даних за використання блочного пошуку / *А. В. Мельничин, Г. Г. Цегелик* // Матеріали всеукр. наук.-практ. конф. "Сучасні інформаційні технології в економіці, менеджменті та освіті". – Львів, 2010. – С. 88–94.
3. *Цегелик Г.Г.* Системы распределенных баз данных / *Г. Г. Цегелик.* – Львов, 1990.
4. *Цегелик Г. Г.* Организация и поиск информации в базах данных / *Г. Г. Цегелик.* – Львов, 1987.

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ОПТИМАЛЬНЫХ СХЕМ ПОИСКА ИНФОРМАЦИИ В ПОСЛЕДОВАТЕЛЬНЫХ ФАЙЛАХ БАЗ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ МЕТОДА БЛОЧНОГО ПОИСКА

А. Мельничин, Г. Цегелик

Львовский национальный университет имени Ивана Франко,
ул. Университетская, 1, Львов, 79000, e-mail: kafmmsep@franko.lviv.ua

Для разных законов распределения вероятностей обращения к записям проведен сравнительный анализ оптимальных схем трех вариантов поиска информации в последовательных файлах баз данных. Как критерий оптимальности принято математическое ожидание общего времени, необходимого для поиска записи в файле.

Ключевые слова: последовательные файлы баз данных, блочный поиск.

COMPARATIVE ANALYSIS OF OPTIMAL SCHEMS OF INFORMATION SEARCH IN SEQUENTIAL DATABASE FILES WITH USING THE BLOCK SEARCH METHOD

A. Melnychyn, H. Tsegelyk

Ivan Franko National University of Lviv,
Universytetska str, 1, Lviv, 79000, e-mail: kafmmsep@franko.lviv.ua

For different laws of distribution probability of request to records carried comparative analysis of optimal schemas of three options search of information in sequential database files.

The mathematical expectation of the total time to search records in a file has been accepted as a optimum criterion.

Key words: sequential database files, block search.

Стаття надійшла до редколегії 14.09.2010
Прийнята до друку 26.01.2011