УДК 004.632.3

А. В. Мельничин, Г. Г. Цегелик (Львівський нац. ун-т імені Івана Франка)

ОПТИМАЛЬНІ СТРАТЕГІЇ ПОШУКУ ЗАПИСІВ В ПОСЛІДОВНИХ ФАЙЛАХ БАЗ ДАНИХ ПРИ ВИКОРИСТАННІ МЕТОДУ БЛОЧНОГО ПОШУКУ В ЛОКАЛІЗОВАНОМУ БЛОЦІ ЗАПИСІВ

For different laws of distribution probability of request to records the efficiency of search in sequential files at using a block search method in the localized block of records, has been investigated

Для різних законів розподілу ймовірностей звертання до записів досліджено ефективність пошуку записів у послідовних файлах у випадку використання методу блочного пошуку в локалізованому блоці записів

Вступ. У [1, 2] для різних законів розподілу ймовірностей звертання до записів [3,4] (рівномірного, "бінарного" , Зіпфа, узагальненого, частковим випадком якого є розподіл, що наближено задовольняє правило "80-20") досліджена ефективність таких двох варіантів пошуку записів у послідовних файлах баз даних з використанням методу послідовного перегляду:

• послідовне читання блоків записів в основну пам'ять і їх послідовний перегляд;

• послідовний перегляд блоку записів, який попередньо локалізований шляхом читання блоків записів в основну пам'ять і перегляду їх останніх записів.

Оскільки послідовний перегляд є найповільнішим методом пошуку, то постає задача дослідження ефективності варіантів пошуку у випадку використання більш швидких методів, зокрема, методу блочного пошуку, методу двійкового пошуку та інших [1, 5].

У роботі для різних законів розподілу ймовірностей звертання до записів проведемо дослідження ефективності пошуку записів у послідовних файлах у випадку використання методу блочного пошуку в блоці записів, який попередньо локалізований шляхом читання блоків записів в основну пам'ять і перегляду їх останніх записів.

Постановка задачі. Припустимо, що файл, який містить N записів, розбитий на n блоків по ms записів в кожному і процес пошуку запису відбувається так. Спочатку локалізується блок, який містить шуканий запис, шляхом послідовного читання блоків записів в основну пам'ять і перегляду їх останніх записів. Після цього пошук потрібного запису продовжується в локалізованому блоці за допомогою методу блочного пошуку. При цьому локалізований блок записів умовно розбивається на s підблоків по m записів в кожному. Представимо математичне сподівання загального часу, необхідного для пошуку запису у файлі, у вигляді суми математичного сподівання часу, необхідного для локалізації блоку записів, математичного сподівання часу, необхідного для локалізації підблоку записів, і математичного сподівання часу, необхідного для пошуку запису в локалізованому підблоці. Тоді математичне сподівання загального часу, необхідного для пошуку запису у файлі, виразиться формулою

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

або

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

де $a = b + dms$ – час читання блоку записів в основну пам'ять, b і d – деякі сталі, t – час перегляду запису в основній пам'яті.

Знайдемо явний вираз для E у випадку різних законів розподілу ймовірностей звертання до записів і визначимо значення параметрів n, s і m, за яких математичне сподівання досягає мінімуму.

Розв'язання задачі. Введемо позначення

$$E_0 = \sum_{i=1}^{n} \sum_{k=1}^{s} \sum_{j=1}^{m} j p_{(i-1)ms+(k-1)m+j}\ , \qquad E_1 = \sum_{i=1}^{n} \sum_{k=1}^{s} \sum_{j=1}^{m} k p_{(i-1)ms+(k-1)m+j}\ ,$$

$$E_2 = \sum_{i=1}^{n} \sum_{k=1}^{s} \sum_{j=1}^{m} i p_{(i-1)ms+(k-1)m+j}\ .$$

Якщо розподіл імовірностей звертання до записів є рівномірним, тобто

$$p_i \ = \ 1/N, \quad i = 1, 2, ..., N,$$

то

$$E_0 = \frac{1}{2}(m+1),\ E_1 = \frac{1}{2}(s+1),\ E_2 = \frac{1}{2}(n+1).$$

Тоді

$$E = \frac{1}{2}\left((n+1)a + (n+s+m+3)t\right),$$

або

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

Оскільки

$$\frac{\partial E}{\partial s} = \frac{1}{2}\left(-\frac{N}{ms^2}(b+t) + dm + t\right),$$

$$\frac{\partial E}{\partial m} = \frac{1}{2}\left(-\frac{N}{m^2 s}(b+t) + ds + t\right),$$

то для визначення параметрів s і m, за яких функція E досягає мінімуму, одержуємо систему рівнянь

$$\begin{cases} ms^2\left(m + \dfrac{t}{d}\right) = \dfrac{b+t}{d}N, \\ m^2 s\left(s + \dfrac{t}{d}\right) = \dfrac{b+t}{d}N, \end{cases}$$

або

$$\begin{cases} s = m, \\ ms^2\left(m + \dfrac{t}{d}\right) = \dfrac{b+t}{d}N. \end{cases}$$

Отже, у випадку рівномірного розподілу ймовірностей звертання до записів $s = m$, де m - додатний корінь рівняння

$$m^3\left(m + \frac{t}{d}\right) = \frac{b+t}{d}N\ .$$

Нехай ймовірності звертання до записів задовольняють "бінарний" розподіл, тобто

$$p_i = 1/2^N, \quad i = 1, 2, ..., N-1, \quad p_N = 1/2^{N-1}.$$

Оскільки [2]

$$E_0 = \frac{m}{2^N} + \left(2 - \frac{m+2}{2^m}\right)\frac{2^m}{2^m - 1}\left(1 - 2^{-N}\right),$$

$$E_1 = \frac{s}{2^N} + \left(\frac{2^m}{2^m - 1} - \frac{s}{2^{ms} - 1}\right)\left(1 - 2^{-N}\right),$$

$$E_2 = \frac{2^{ms}}{2^{ms} - 1}\left(1 - 2^{-N}\right),$$

то, нехтуючи величиною 2^{-N}, з достатньо високою точністю можемо прийняти

$$E = \frac{2^{ms}}{2^{ms} - 1}a + \left(\frac{2^{ms}}{2^{ms} - 1} + \frac{2^m}{2^m - 1} - \frac{s}{2^{ms} - 1} + \left(2 - \frac{m+2}{2^m}\right)\frac{2^m}{2^m - 1}\right)t,$$

або

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

Звідси одержуємо такий вираз для E

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

Якщо E розділити на d, то матимемо

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

Оскільки $b/d = const$, $t/d = const$, то в області $m \geq 2$, $s \geq 2$, функція E/d приймає найменше значення при $m = s = 2$. Отже, в області $m \geq 2$, $s \geq 2$

$$\min E/d = \frac{16}{15}\left(\frac{b}{d} + 4\right) + \frac{18}{5} \cdot \frac{t}{d}.$$

Припустимо, що ймовірності звертання до записів розподілені за законом Зіпфа, тобто

$$p_i = \frac{1}{iH_N}, \quad i = 1, 2, ..., N,$$

де $H_N = \sum_{k=1}^{N} 1/k$ – частинна сума гармонічного ряду. Оскільки [2]

$$E_0 = \frac{1}{H_N}\left(m \cdot S_m(sn) - (H_N - 1)N\right),$$

$$E_1 = \frac{1}{H_N}\left(H_N + s \cdot S_{ms}(n) - S_m(sn)\right),$$

$$E_2 = \frac{1}{H_N}\left((n+1)H_N - S_{ms}(n)\right),$$

то

$$E = \frac{1}{H_N}\left(\left((n+1)H_N - S_{ms}(n)\right)a + \left((n+2)H_N + \right.\right.$$
$$\left.\left. + (s-1)S_{ms}(n) + (m-1)S_m(sn) - (H_N - 1)N\ \right)t\right)\ ,$$

де

$$S_{ms}(n) = \sum_{k=1}^{n} H_{kms}, \quad S_m(sn) = \sum_{k=1}^{sn} H_{km}.$$

Використовуючи апроксимацію $S_{ms}(n)$ і $S_m(sn)$ відповідно виразами [4]

$$\overline{S}_{ms}(n) = n(H_N - 1) + \frac{1}{2}\ln n + C_1,$$

$$\overline{S}_m(sn) = sn(H_N - 1) + \frac{1}{2}\ln sn + C_1,$$

де $C_1 = \frac{1}{2}\ln 2\pi$, з достатньо високою точністю можемо прийняти

$$E = \frac{1}{H_N}\left[\left(H_N + n - \frac{1}{2}\ln n - C_1\right)a + \right.$$
$$\left.\left(2H_N + n + \frac{1}{2}((s+m-2)\ln n + (m-1)\ln s) + (s+m-2)C_1\right)t\right]\ ,$$

або

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

Оскільки

$$\frac{\partial E}{\partial s} = \frac{1}{H_N}\left(\frac{1}{2}(dm - t)\left(1 - \ln\frac{N}{ms} - 2C_1\right) + \frac{1}{2s}(b+t) + H_N dm - \frac{N}{ms^2}(b+t)\right)\ ,$$
$$\frac{\partial E}{\partial m} = \frac{1}{H_N}\left(\frac{1}{2}(ds - t)\left(1 - \ln\frac{N}{ms} - 2C_1\right) + \frac{1}{2m}(b - st) + H_N ds - \frac{N}{m^2 s}(b+t) + \right.$$
$$\left. + \left(\frac{1}{m} + \frac{\ln s}{2}\right)t\right),$$

то для знаходження параметрів s і m, за яких E досягає мінімуму, одержуємо систему рівнянь

$$\begin{cases} ms^2\left(m-\dfrac{t}{d}\right)\left(1-\ln\dfrac{N}{ms}-2C_1\right)+ms\dfrac{b+t}{d}+2m^2s^2H_N=2N\dfrac{b+t}{d}, \\ m^2s\left(s-\dfrac{t}{d}\right)\left(1-\ln\dfrac{N}{ms}-2C_1\right)+ms\left(\dfrac{b-st}{d}\right)+2m^2s^2H_N+ \\ +2m^2s\left(\dfrac{1}{m}+\dfrac{\ln s}{2}\right)\dfrac{t}{d}=2N\left(\dfrac{b+t}{d}\right). \end{cases}$$

Якщо відняти від першого рівняння друге, то одержимо систему

$$\begin{cases} 2m^2s^2H_N-ms^2\left(m-\dfrac{t}{d}\right)\left(\ln\dfrac{N}{ms}+2C_1-1\right)+ms\left(\dfrac{b+t}{d}\right)=2N\left(\dfrac{b+t}{d}\right) \\ (s-m)\left(\ln N+2C_1\right)-s\ln ms+m(1+\ln m)-1=0\,, \end{cases},$$

або

$$\begin{cases} ms\left(2msH_N-s\left(m-\dfrac{t}{d}\right)\left(\ln N-\ln ms+2C_1-1\right)+\dfrac{b+t}{d}\right)=2N\left(\dfrac{b+t}{d}\right), \\ (s-m)\left(\ln N-\ln m+2C_1\right)-s\ln s+m-1=0. \end{cases}$$

Якщо ймовірності звертання до записів задовольняють узагальнений закон розподілу, тобто

$$p_i=\frac{1}{i^cH_N^{(c)}},\quad i=1,2,...,N,$$

де c – будь-який параметр $(0<c<1)$, $H_N^{(c)}=\sum_{k=1}^{N}1/k^c$ – частинна сума узагальненого гармонічного ряду, то [2]

$$E_0=\frac{1}{H_N^{(c)}}\left(m\cdot S_m^{(c)}(sn)+H_N^{(c-1)}-NH_N^{(c)}\right),$$

$$E_1=\frac{1}{H_N^{(c)}}\left(s\cdot S_{ms}^{(c)}(n)+H_N^{(c)}-S_m^{(c)}(sn)\right),$$

$$E_2=\frac{1}{H_N^{(c)}}\left((n+1)H_N^{(c)}-S_{ms}^{(c)}(n)\right).$$

Тоді

$$E=\frac{1}{H_N^{(c)}}\left(\left((n+1)H_N^{(c)}-S_{ms}^{(c)}(n)\right)a+\right.$$
$$\left.+\left(H_N^{(c-1)}+(n+2-N)H_N^{(c)}+(s-1)S_{ms}^{(c)}(n)+(m-1)S_m^{(c)}(sn)\right)t\right)\ ,$$

де

$$S_{ms}^{(c)}(n)=\sum_{k=1}^{n}H_{kms}^{(c)},\quad S_m^{(c)}(sn)=\sum_{k=1}^{sn}H_{km}^{(c)}.$$

Використовуючи апроксимацію $S_{ms}^{(c)}(n)$ і $S_m^{(c)}(sn)$ відповідно виразами [4]

$$\overline{S}_{ms}^{(c)}(n)=nH_N^{(c)}+\frac{N^{1-c}}{1-c}\left(\frac{c-1}{2-c}n+\frac{\alpha^{(c)}(n)}{n^{1-c}}\right),$$

$$\overline{S}_m^{(c)}(sn) = snH_N^{(c)} + \frac{N^{1-c}}{1-c}\left(\frac{c-1}{2-c}sn + \frac{\alpha^{(c)}(sn)}{(sn)^{1-c}}\right),$$

де

$$\alpha^{(c)}(n) = H_n^{(c-1)} - \frac{1}{2-c}n^{2-c},$$

$$\alpha^{(c)}(sn) = H_{sn}^{(c-1)} - \frac{1}{2-c}(sn)^{2-c},$$

з достатньо високою точністю можемо прийняти

$$E = \frac{1}{H_N^{(c)}}\left(\left(H_N^{(c)} - \frac{N^{1-c}}{1-c}\left(\frac{c-1}{2-c}n + \frac{\alpha^{(c)}(n)}{n^{1-c}}\right)\right)a + \right.$$
$$\left. + \left(2H_N^{(c)} + H_N^{(c-1)} + \frac{N^{1-c}}{1-c}\left((N-n)\frac{c-1}{2-c} + \frac{(s-1)\alpha^{(c)}(n)}{n^{1-c}} + \frac{(m-1)\alpha^{(c)}(sn)}{(sn)^{1-c}}\right)\right)t\right),$$

або

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

В табл. 1 наведені значення параметрів n, m, s при яких математичне сподівання досягає мінімуму, для різних законів розподілу ймовірностей звертання до записів, деяких значень b/d, $t/d = 0.1$ і $N = 10^6$.

Таблиця 1.

Значення параметрів оптимальної організації пошуку

b/d	Параметри	*Закон розподілу*						
		$c = 0$	$c = 0.2$	$c = 0.4$	$c = 0.6$	$c = 0.8$	$c = 1$	"Бінарний"
10	n	314.9	334.6	366.4	427.9	578.0	1021.0	297899.2
	m	56,3	54.6	52.0	47.6	39.6	27.5	1.7
	s	56,3	54.7	52.4	49.1	43.7	35.6	1.9
100	n	100.0	106.5	117.2	138.1	188.4	333.1	160783.2
	m	100.0	96.7	91.6	83.1	68.9	46.7	2.0
	s	100.0	97.1	93.1	87.2	77.0	64.3	3.1
1000	n	31.6	33.9	37.6	44.8	61.5	108.1	105804.5
	m	177.8	171.0	160.6	143.8	115.3	77.8	2.1
	s	177.8	172.7	165.7	155.4	141.1	118.9	4.5

В табл. 1 значенню $c = 0$ відповідає рівномірний розподіл, значенню $c = 1$ — закон Зіпфа.

На мал. 1 показана залежність оптимального значення величини E/d від зміни закону розподілу ймовірностей звертання до записів для деяких b/d, $t/d = 0.1$ і $N = 10^6$. Із мал. 1 видно, що зі зміною закону розподілу ймовірностей звертання до записів суттєво змінюється оптимальне значення величини E/d.

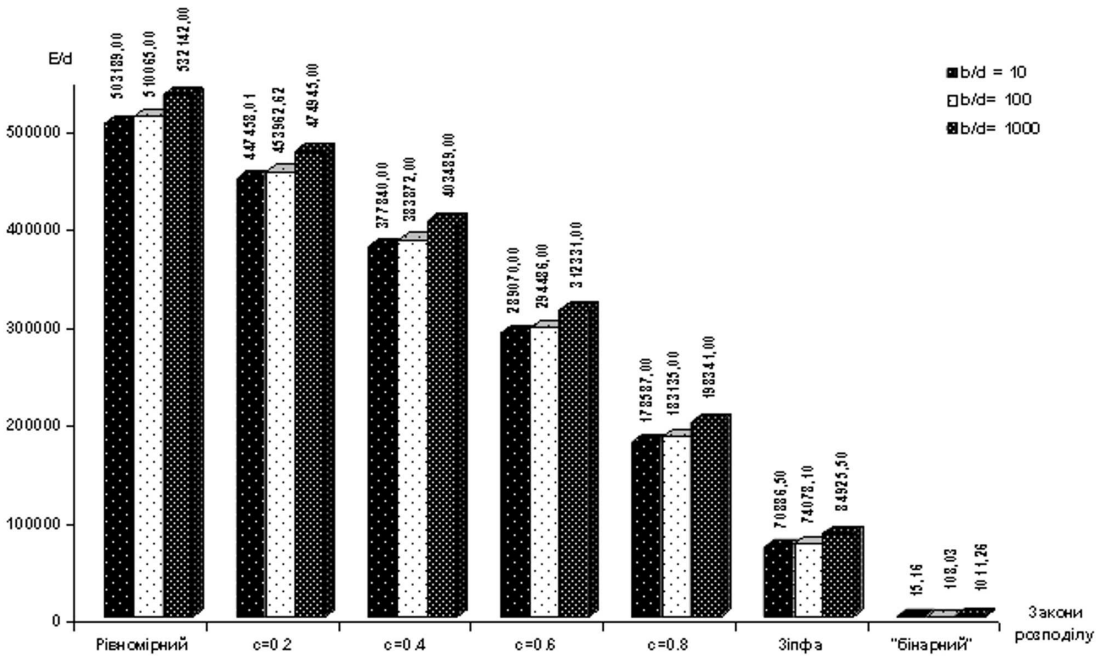

Рис. 1. Оптимальне значення величини E/d для різних законів розподілу ймовірностей звертання до записів для деяких b/d, $t/d = 0.1$ і $N = 10^6$

Висновок. Проведено дослідження ефективності пошуку записів у послідовних файлах у випадку використання методу блочного пошуку в блоці записів, який попередньо локалізований шляхом читання блоків записів в основну пам'ять і перегляду їх останніх записів, для різних законів розподілу ймовірностей звертання до записів. Для кожного закону розподілу ймовірностей звертання до записів виведені співвідношення для визначення параметрів оптимальної організації пошуку. Проведений розрахунок оптимальних параметрів для розглянутих законів розподілу ймовірностей звертання до записів для декількох конкретних випадків.

1. *Кудеравець Х.С., Цегелик Г.Г.* Побудова та аналіз оптимальних стратегій пошуку записів в послідовних файлах для різних законів розподілу ймовірностей звертання до записів. — Львів: 1997. — 70с. (Препринт / Львівський державний університет ім. І. Франка; №1 – 97).
2. *Цегелик Г.Г.* Системы распределенных баз данных. — Львов: Свит, 1990. — 168 с.
3. *Кнут Д.* Искусство программирования для ЭВМ. Т. 3: Сортировка и поиск. — М.: Издательский дом "Вильямс", 2000. — 832 с.
4. *Цегелик Г.Г.* Организация и поиск информации в базах данных. — Львов: Вища школа, 1987. — 176 с.
5. *Мартин Дж.* Организация баз данных в вычислительных системах. — М.: Мир, 1980. — 644 с.

Одержано 17.06.2009