УДК 004.632.3

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

ДРУГИЙ ВАРІАНТ ПОБУДОВИ ОПТИМАЛЬНИХ СТРАТЕГІЙ ПОШУКУ ЗАПИСІВ У ПОСЛІДОВНИХ ФАЙЛАХ БАЗ ДАНИХ У ВИПАДКУ ВИКОРИСТАННЯ БЛОЧНОГО ПОШУКУ

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 блоці записів, який попередньо локалізований шляхом читання і перегляду останніх записів кожного блоку

Вступ. У [1, 2] для різних законів розподілу ймовірностей звертання до записів [3] (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, як послідовне читання блоків записів в основну пам'ять і їх послідовний перегляд та 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.

У роботі [4] досліджена ефективність використання методу блочного пошуку в блоці записів, який попередньо локалізований шляхом читання блоків записів в основну пам'ять і перегляду їх останніх записів.

Враховуючи те, що метод блочного пошуку є одним із найуживаніших, то проведемо дослідження ефективності використання цього методу для пошуку записів у блоці, який попередньо локалізований шляхом читання і перегляду останніх записів кожного блоку.

Постановка задачі. Припустимо, що файл, який містить N записів, розбитий на n блоків по ms записів у кожному і процес пошуку запису відбувається так. Спочатку локалізується блок, який містить шуканий запис, шляхом послідовного читання і перегляду останніх записів кожного блоку. 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

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

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 $a = b + dms$ – час читання блоку записів в основну пам'ять, b і d – деякі сталі, $t_1 = b + d + t$ – час читання запису в основну пам'ять і його перегляд, 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, загального часу, необхідного для пошуку запису у файлі, досягає мінімуму.

Розв'язання задачі. Введемо позначення

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

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

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

Якщо розподіл імовірностей звертання до записів є рівномірним, тобто

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

то

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

Тоді

$$E = a + \frac{1}{2}(n+1)t_1 + \frac{1}{2}(s+m+2)t,$$

або

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

На рис. 1 наведена поведінка функції E/d в околі точки мінімумуу у випадку рівномірного розподілу ймовірностей звертання до записів, $b/d = 1000$, $t/d \ = \ 0.1$ і $N = 10^6$.

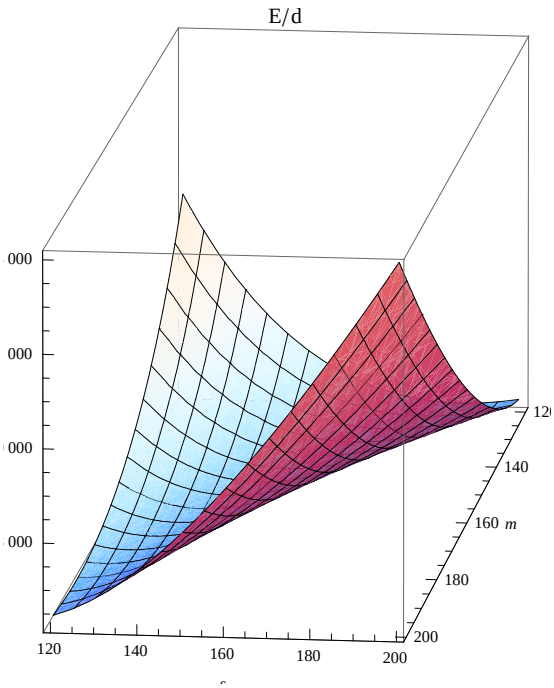

Рис. 1. Поведінка функції E/d в околі точки мінімумуу у випадку рівномірного розподілу ймовірностей звертання до записів, $b/d = 1000$, $t/d = 0.1$ і $N = 10^6$

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

$$E = a + \frac{2^{ms}}{2^{ms} - 1} t_1 + \left(\frac{2^m}{2^m - 1} \left(2 - \frac{m+2}{2^m}\right) + \frac{2^m}{2^m - 1} - \frac{s}{2^{ms} - 1}\right) t,$$

або

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

На рис. 2 показана поведінка функції E/d в околі точки мінімуму у випадку "біннарного" розподілу ймовірностей звертання до записів, $b/d = 1000$, $t/d = 0.1$ і $N = 10^6$.

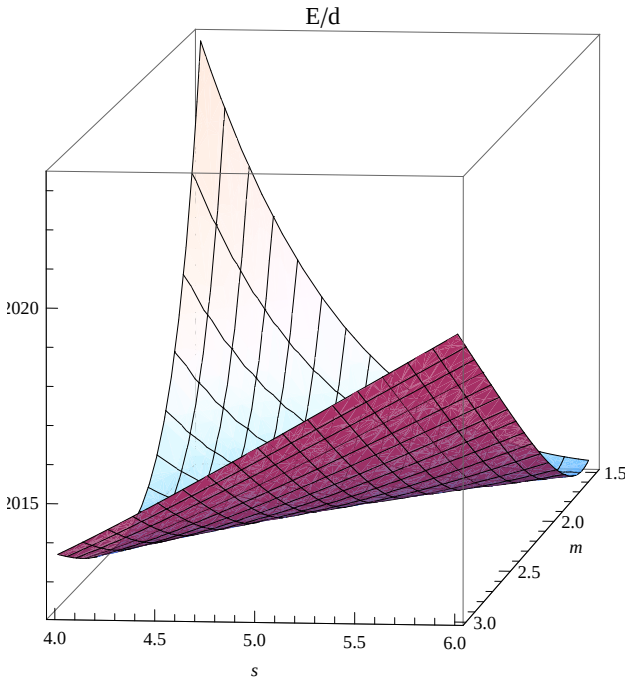

Рис. 2. Поведінка функції E/d в околі точки мінімуму у випадку "біннарного" розподілу ймовірностей звертання до записів, $b/d = 1000$, $t/d = 0.1$ і $N = 10^6$

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

$$E = a + \frac{1}{H_N}\left((n+1)H_N - S_{ms}(n)\right)t_1 + \frac{1}{H_N}\left(H_N + \right.$$
$$\left. + s \cdot S_{ms}(n) - S_m(sn) + m \cdot S_m(sn) - (H_N - 1)N\right)t) \ ,$$

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 [3]

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

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

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

$$E = a + \frac{1}{H_N}\left(\left(H_N + n - \frac{1}{2}\ln n - C_1\right)t_1 + \right.$$
$$\left. + \left(H_N + \frac{1}{2}\left((s+m-1)\left(\ln n + 2C_1\right) + (m-1)\ln s\right)\right)t\right) \ ,$$

або

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

На рис. 3 показана поведінка функції E/d в околі точки мінімуму у випадку розподілу записів за законом Зіпфа, $b/d = 1000$, $t/d = 0.1$ і $N = 10^6$.

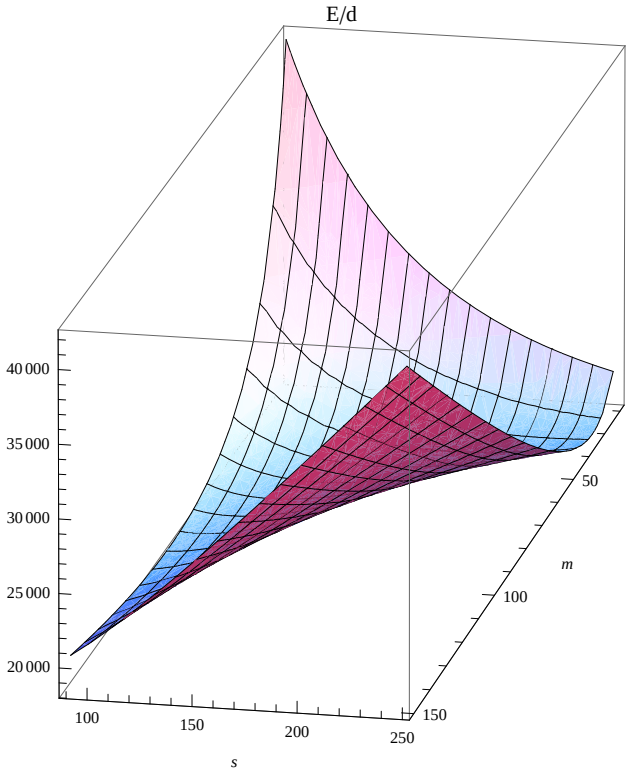

Рис. 3. Поведінка функції E/d в околі точки мінімуму у випадку розподілу записів за законом Зіпфа, $b/d = 1000$, $t/d = 0.1$ і $N = 10^6$

Якщо ймовірності звертання до записів задовольняють узагальнений закон розподілу, тобто

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

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]

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

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

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

Тоді

$$E = a + \frac{1}{H_N^{(c)}} \left(\left((n+1) H_N^{(c)} - S_{ms}^{(c)}(n) \right) t_1 + \left(H_N^{(c-1)} + (1-N) H_N^{(c)} + \right. \right.$$

$$\left. \left. + \, s \cdot S_{ms}^{(c)}(n) + (m-1) S_m^{(c)}(sn) \right) t \right) ,$$

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 [3]

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

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

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

$$E = a + \frac{1}{H_N^{(c)}} \left(\left(H_N^{(c)} - \frac{N^{1-c}}{1-c} \left(\frac{c-1}{2-c} n + \frac{\alpha^{(c)}(n)}{n^{1-c}} \right) \right) t_1 + \left(H_N^{(c-1)} + \right. \right.$$

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

або

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

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

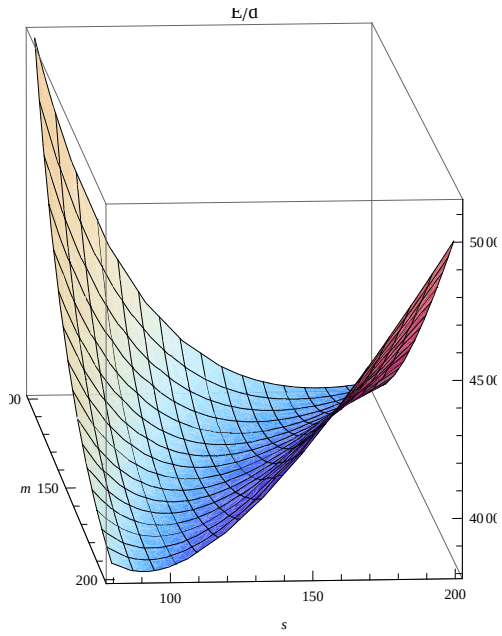

Рис. 4. Поведінка функції E/d в околі точки мінімуму у випадку узагальненого розподілу ймовірностей звертання до записів, c=0.5, $b/d = 1000$, $t/d = 0.1$ і $N = 10^6$

На рис. 4 показана поведінка функції E/d в околі точки мінімуму у випадку узагальненого розподілу ймовірностей звертання до записів, c=0.5, $b/d = 1000$, $t/d = 0.1$ і $N = 10^6$.

У табл. 1 для різних законів розподілу ймовірностей звертання до записів і для деяких значень b/d, $t/d = 0.1$ та $N = 10^6$, наведені оптимальні значення параметрів n, m та s (обчислені із використанням функцій пакету Wolfram Mathematica 7.0), за яких математичне сподівання загального часу, необхідного для пошуку запису у файлі, досягає мінімуму.

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"
10^1	n	425	450	490	561	715	1140	307862
	m	49	47	45	42	36	26	3
	s	49	47	45	43	39	34	1
10^2	n	141	149	162	186	237	378	162035
	m	84	82	78	72	61	43	2
	s	84	82	79	75	70	61	2
10^3	n	445	47	52	59	75	120	105904
	m	150	145	137	125	104	73	2
	s	150	146	141	135	127	114	4

У табл. 2 (з точністю до 0.1) наведені оптимальні значення величини E/d

для різних законів розподілу ймовірностей звертання до записів для деяких b/d і N, $t/d = 0.1$.

Таблиця 2.

Оптимальні значення величини E/d

N	b/d	*Закон розподілу*						
		$c = 0$	$c = 0.2$	$c = 0.4$	$c = 0.6$	$c = 0.8$	$c = 1$	"Бінарний"
10^4	10	488, 4	461, 5	425, 8	377, 1	311, 2	231, 3	25, 9
	100	1575, 3	1495, 0	1389, 0	1245, 1	1051, 6	817, 9	209, 0
	1000	5980, 0	5745, 4	5439, 4	5028, 7	4479, 7	3813, 9	2012, 3
10^5	10	1508, 4	1423, 1	1309, 0	1149, 0	919, 6	625, 0	25, 9
	100	4652, 1	4395, 3	4052, 1	3571, 5	2884, 9	2004, 8	209, 0
	1000	15660, 0	14859, 3	13795, 9	12317, 3	10218, 8	7536, 3	2012, 3
10^6	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

У табл. 1, 2 значенню $c = 0$ відповідає рівномірний розподіл, значенню $c = 1$ — закон Зіпфа.

На рис. 5 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 рис. 5 видно, що зі зміною закону розподілу ймовірностей звертання до записів суттєво змінюється оптимальне значення величини E/d.

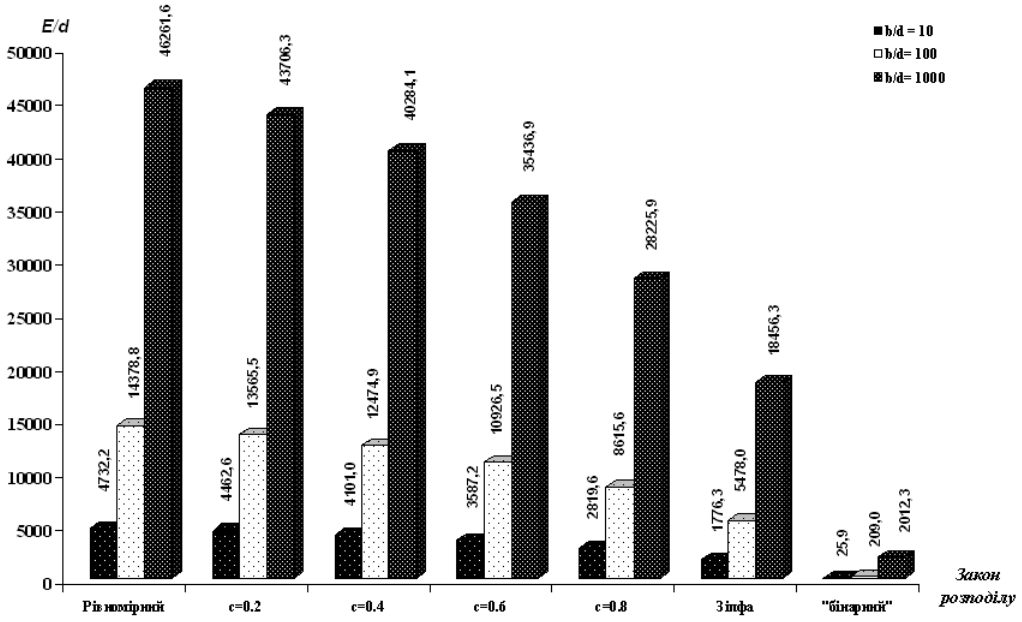

Рис. 5. Оптимальне значення величини E/d для різних законів розподілу ймовірностей звертання до записів для деяких b/d, $t/d = 0.1$ і $N = 10^6$

Висновки. Проведено дослідження ефективності пошуку записів у послідовних файлах за використання методу блочного пошуку в блоці записів, який попередньо локалізований шляхом читання і перегляду останніх записів кожного блоку, 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.* Побудова та аналіз оптимальних стратегій пошуку записів в послідовних файлах для різних законів розподілу ймовірностей звертання до записів / *Кудеравець Х.С., Цегелик Г.Г.* — Львів: 1997. — 70с. (Препринт / Львівський державний університет ім. І. Франка; №1 – 97).
2. *Цегелик Г.Г.* Системы распределенных баз данных / *Г.Г. Цегелик* — Львов: Свит, 1990. — 168 с.
3. *Цегелик Г.Г.* Организация и поиск информации в базах данных / *Г.Г. Цегелик* — Львов: Вища школа, 1987. — 176 с.
4. *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.

Одержано 29.04.2011