УДК 004.632

М. І. Глебена (ДВНЗ «Ужгородський нац. ун-т»),
В. Ф. Глебена, О. М. Попельський (Закарпатський науково-дослідний експертно-криміналістичний центр МВС України)

ВИЗНАЧЕННЯ ОПТИМАЛЬНИХ ПАРАМЕТРІВ МОДЕЛЕЙ ДОСТУПУ ДО ІНФОРМАЦІЇ У ФАЙЛАХ БАЗ ДАНИХ.

The paper considers models of optimal organization of sequential database files. The case of a generalized law of distribution of probabilities of access to records is considered. To determine the optimal parameters of the model, methods of majorant type are used.

У роботі розглядаються моделі оптимальної організації послідовних файлів баз даних. Розглянуто випадок узагальненого закону розподілу ймовірностей звертання до записів. Для визначення оптимальних параметрів моделі використано методи мажорантного типу.

1. Вступ. Головною тенденцією розвитку сучасної індустрії інформатики є створення обчислювальних систем, здатних опрацьовувати величезні обсяги інформації в режимі реального або мінімального масштабу часу. Головна концепція проектування таких систем — забезпечення їхньої високої продуктивності. Одним із напрямів реалізації вказаної концепції є удосконалення технології опрацювання інформації в обчислювальних системах. Оскільки основу сучасних інформаційних технологій складають бази даних (БД) і системи керування базами даних (СКБД), то удосконалення технології опрацювання інформації з використанням концепції баз даних передбачає в першу чергу розв'язання проблеми оптимальної організації та пошуку інформації у файлах баз даних. Така організація забезпечує доступ користувачів до інформації БД за мінімально допустимий час і в значній мірі визначається ефективністю методів пошуку інформації у файлах БД.

У більшості систем опрацювання інформації типовими є випадки нерівномірного розподілу ймовірностей звертання до записів. У роботі розглянуто моделі оптимального доступу до інформації файлів баз даних у випадку узагальненого закону розподілу ймовірностей звертання до записів. Для відшукання оптимальних параметрів моделі використано методи мажорантного типу оптимізації функцій однієї та двох дійсних змінних [1, 2].

2. Блоковий пошук з оптимальним розміром блоків. Якщо записи файла впорядковані за зростанням чи спаданням значень ключа, то для пошуку запису не обов'язково переглядати всі записи, що передують шуканому. Записи файла можна розбити на блоки і спочатку локалізувати блок, який містить шуканий запис, переглядаючи останні записи блоків. Після того, як блок записів локалізований, пошук потрібного запису в цьому блоці можна продовжити за допомогою методу послідовного перегляду [3, 4].

Нехай усі записи впорядкованого файла розбиті на n блоків по m записів у кожному ($N{=}nm$) і p_i – ймовірність звертання до i-го запису файла. За критерій ефективності приймемо математичне сподівання кількості порівнянь, необхідних для пошуку запису у файлі. Математичне сподівання кількості порівнянь запишемо у вигляді суми математичного сподівання кількості порівнянь, необхідних для локалізації блока, який містить шуканий запис, і математичного сподівання кількості порівнянь, необхідних для пошуку запису в локалізованому блоці. Тоді математичне сподівання кількості порівнянь, необхідних для пошуку запису у файлі, виражається формулою

$$E = \sum_{i=1}^{n} i \left(\sum_{j=1}^{m} p_{(i-1)m+j} \right) + \sum_{i=1}^{n} \sum_{j=1}^{m} j p_{(i-1)m+j}. \tag{1}$$

або

$$E = \sum_{i=1}^{n} \sum_{j=1}^{m} (i+j)\, p_{(i-1)m+j}. \tag{2}$$

Запишемо вираз для E у випадку узагальненого розподілу ймовірностей звертання до записів. Припустимо, що ймовірності звертання до записів задовольняють узагальнений закон розподілу. Тоді, для E одержимо вираз

$$E = \frac{1}{H_N^{(c)}} \left(H_N^{(c-1)} - (N-n-1)\, H_N^{(c)} + (m-1)\, S_m^{(c)}(n) \right), \tag{3}$$

де

$$S_m^{(c)}(n) = \sum_{i=1}^{n} H_{im}^{(c)}.$$

Використовуючи апроксимацію $S_m^{(c)}(n)$ виразом

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

де $\alpha^{(c)}(n) = H_n^{(c-1)} - \frac{1}{2-c} n^{2-c}$; $(0 \le c \le 1)$ — повільно зростаюча функція, а $H_n^{(c-1)} = \sum_{k=1}^{n} \frac{1}{k^{c-1}}$; $H_N^{(c)} = \sum_{k=1}^{N} \frac{1}{k^c}$, тоді з достатньо високою точністю можемо прийняти

$$E = \frac{1}{H_N^{(c)}} \left(H_N^{(c-1)} + H_N^{(c)} + \frac{(m-1)\, N^{1-c}}{1-c} \left(\frac{c-1}{2-c} n + \frac{\alpha^{(c)}(n)}{n^{1-c}} \right) \right), \tag{4}$$

або

$$E = \frac{1}{H_N^{(c)}} \left(H_N^{(c-1)} + H_N^{(c)} + \left(\frac{N}{n} - 1 \right) \frac{N^{1-c}}{1-c} \left(\frac{c-1}{2-c} n + \frac{\alpha^{(c)}(n)}{n^{1-c}} \right) \right). \tag{5}$$

Модель визначення параметрів оптимального блокового пошуку для узагальненого розподілу ймовірностей звертання до записів можемо записати у вигляді

$$E = \frac{1}{H_N^{(c)}} \left(H_N^{(c-1)} + H_N^{(c)} + \left(\frac{N}{n} - 1 \right) \frac{N^{1-c}}{1-c} \left(\frac{c-1}{2-c} n + \frac{\alpha^{(c)}(n)}{n^{1-c}} \right) \right) \to \min. \tag{6}$$

Така модель блокового пошуку називається блоковим пошуком з оптимальним розміром блоків.

Функція E є опуклою, тоді для знаходження значення параметра n, за якого функція досягає мінімуму, використаємо алгоритм відшукання екстремуму для

Таблиця 1

N	***c =0.2***	***c =0.4***	***c =0.6***	***c =0.8***
10^3	33	35	38	43
10^4	106	114	126	144
10^5	335	362	405	476
10^6	1060	1150	1297	1556

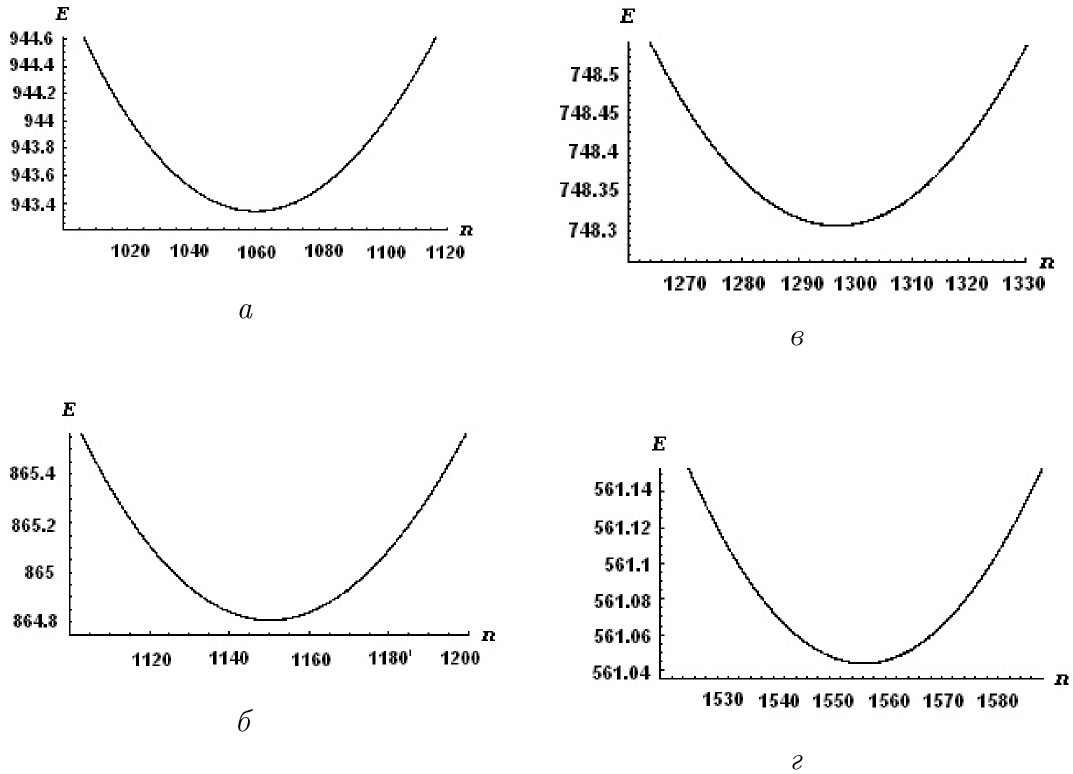

Рис. 1. Графік поведінки математичного сподівання E в околі точки мінімуму для $N = 10^6$ і *c = 0.2(а), c = 0.4(б), c = 0.6(в), c = 0.8(г)*

логарифмічно вгнутих функцій [3]. Тобто застосуємо метод мажорантного типу до задачі $-E\left(n\right) \to \max$.

Значення оптимального параметру n для різних c та N наведено в таблиці 1.

Графік поведінки математичного сподівання E в околі точки мінімуму для $N = 10^6$ і різних значень параметра c зображено на рис. 1.

3. Дворівневий блоковий пошук з оптимальним розміром блоків і підблоків. У випадку, коли всі записи файла розбиті на n блоків по m записів у кожному, а кожний блок записів, відповідно, – на l підблоків по s записів у кожному пошук запису у файлі відбувається так: спочатку локалізуємо блок, який містить шуканий запис, шляхом перегляду останніх записів блоків. Після цього в локалізованому блоці шукаємо підблок, який містить шуканий запис, шляхом перегляду останніх записів підблоків. І, нарешті, у локалізованому підблоці запис шукаємо методом послідовного перегляду.

Нехай p_i – імовірність звертання до i-го запису файла. Запишемо математичне сподівання кількості порівнянь, необхідних для пошуку запису у файлі, у вигляді суми: математичного сподівання кількості порівнянь, необхідних для локалізації блока, математичного сподівання кількості порівнянь, необхідних для локалізації підблока, і математичного сподівання кількості порівнянь, необхідних для пошуку запису у локалізованому підблоці. Тоді математичне сподівання кількості порівнянь, необхідних для пошуку запису у файлі, виражається формулою

$$E = \textstyle\sum_{i=1}^{n}\left(i\sum_{j=1}^{l}\sum_{k=1}^{s}p_{(i-1)ls+(j-1)s+k}\right)+\sum_{i=1}^{n}\sum_{j=1}^{l}\left(j\sum_{k=1}^{s}p_{(i-1)ls+(j-1)s+k}\right)+ \\ +\sum_{i=1}^{n}\sum_{j=1}^{l}\sum_{k=1}^{s}kp_{(i-1)ls+(j-1)s+k}, \tag{7}$$

або

$$E=\sum_{i=1}^{n}\sum_{j=1}^{l}\sum_{k=1}^{s}(i+j+k)\,p_{(i-1)ls+(j-1)s+k}\ . \tag{8}$$

Запишемо вираз для E у випадку узагальненого розподілу ймовірностей звертання до записів. Нехай ймовірності звертання до записів задовольняють узагальнений закон розподілу. Тоді, для E одержимо вираз

$$E=\tfrac{1}{H_N^{(c)}}\left((n+1)\,H_N^{(c)}-S_{sl}^{(c)}(n)\right)+\tfrac{1}{H_N^{(c)}}\left(H_N^{(c)}+l\cdot S_{sl}^{(c)}(n)-S_s^{(c)}(nl)\right)+ \\ +\tfrac{1}{H_N^{(c)}}\left(s\cdot S_s^{(c)}(nl)+H_N^{(c-1)}-NH_N^{(c)}\right), \tag{9}$$

де

$$S_{sl}^{(c)}(n)=\sum_{k=1}^{n}H_{ksl}^{(c)},\qquad S_s^{(c)}(nl)=\sum_{k=1}^{nl}H_{ks}^{(c)}.$$

Використаємо апроксимації $S_{sl}^{(c)}(n)$ і $S_s^{(c)}(nl)$, відповідно, виразами

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

$$\bar{S}_{s}^{(c)}(nl)=nlH_N^{(c)}+\frac{N^{1-c}}{1-c}\left(\frac{c-1}{2-c}nl+\frac{\alpha^{(c)}(nl)}{(nl)^{1-c}}\right),$$

де

$$\alpha^{(c)}(n)=H_n^{(c-1)}-\frac{1}{2-c}n^{2-c},(0\le c\le 1),$$

$$\alpha^{(c)}(nl)=H_{nl}^{(c-1)}-\frac{1}{2-c}(nl)^{2-c},(0\le c\le 1),$$

$$H_N^{(c)}=\sum_{k=1}^{N}\frac{1}{k^c},\qquad H_n^{(c-1)}=\sum_{k=1}^{n}\frac{1}{k^{c-1}},\qquad H_{nl}^{(c-1)}=\sum_{k=1}^{nl}\frac{1}{k^{c-1}}.$$

Тоді з достатньо високою точністю можемо прийняти

$$E=\tfrac{1}{H_N^{(c)}}\left(H_N^{(c)}-\tfrac{N^{1-c}}{1-c}\left(\tfrac{c-1}{2-c}n+\tfrac{\alpha^{(c)}(n)}{n^{1-c}}\right)\right)+ \\ +\tfrac{1}{H_N^{(c)}}\left(H_N^{(c)}-\tfrac{N^{1-c}}{1-c}\left(\tfrac{l\cdot\alpha^{(c)}(n)}{n^{1-c}}-\tfrac{\alpha^{(c)}(nl)}{(nl)^{1-c}}\right)\right)+ \\ +\tfrac{1}{H_N^{(c)}}\left(H_N^{(c-1)}-\tfrac{N^{2-c}}{1-c}\left(\tfrac{c-1}{2-c}+\tfrac{\alpha^{(c)}(nl)}{(nl)^{2-c}}\right)\right), \tag{10}$$

або

$$E = \frac{1}{H_N^{(c)}} \left(H_N^{(c-1)} + 2H_N^{(c)} - \frac{N^{1-c}}{1-c} \left(\frac{c-1}{2-c} n - (l-1) \frac{\alpha^{(c)}(n)}{n^{1-c}} + \frac{\alpha^{(c)}(nl)}{(nl)^{1-c}} \right) + \right.$$
$$\left. + \frac{N^{2-c}}{1-c} \left(\frac{c-1}{2-c} + \frac{\alpha^{(c)}(nl)}{(nl)^{2-c}} \right) \right). \tag{11}$$

Тоді модель оптимального дворівневого блокового пошуку для узагальненого закону розподілу ймовірностей звертання до записів, запишемо у вигляді

$$E = \frac{1}{H_N^{(c)}} \left(H_N^{(c-1)} + 2H_N^{(c)} - \frac{N^{1-c}}{1-c} \left(\frac{c-1}{2-c} n - (l-1) \frac{\alpha^{(c)}(n)}{n^{1-c}} + \frac{\alpha^{(c)}(nl)}{(nl)^{1-c}} \right) + \right.$$
$$\left. + \frac{N^{2-c}}{1-c} \left(\frac{c-1}{2-c} + \frac{\alpha^{(c)}(nl)}{(nl)^{2-c}} \right) \right) \to \min. \tag{12}$$

Така модель називається дворівневим блоковим пошуком з оптимальним розміром блоків і підблоків.

Оскільки функція E у виразі (12) є опуклою, то оптимальні значення параметрів n та l, за яких E досягає мінімуму, будемо шукати, використовуючи алгоритм покоординатного підйому відшукання екстремуму логарифмічно вгнутих функцій двох дійсних змінних [4]. Тобто одержимо задачу $-E(n,l) \to \max$. Оптимальні значення математичного сподівання кількості порівнянь, необхідних для пошуку запису у файлі, для різних значень параметра c та N наведено в таблиці 2.

Таблиця 2

N	c	n_{on}	l_{on}	s_{on}	E_{on}
100000	0.2	50	45	44.44	68.16
	0.4	54	44	42.09	63.92
	0.6	62	42	38.40	57.22
	0.8	77	39	33.30	46.15
500000	0.2	86	76	76.49	115.66
	0.4	95	73	72.09	108.66
	0.6	110	69	65.87	97.46
	0.8	135	66	56.11	77.89
1000000	0.2	108	96	96.45	145.40
	0.4	120	92	90.57	136.69
	0.6	139	87	82.69	122.73
	0.8	173	82	70.49	97.85

3. Висновок. У роботі розглянуто побудову моделей оптимального доступу до інформації файлів баз даних у випадку узагальненого закону розподілу ймовірностей звертання до записів. Використовуючи методи мажорантного типу знайдено оптимальні параметри моделі при різних значеннях N та c.

Список використаної літератури

1. *Глебена М. І.* Модифікований чисельний метод відшукання абсолютного екстремуму негладких і розривних функцій. / М.І.Глебена, Г.Г.Цегелик // Наук. вісник Ужгород. ун-ту. Сер. матем. і інформ. – 2008. – Вип. 16. – С. 57-61.

2. *Глебена М. І.* Алгоритм відшукання максимального значення довільної логарифмічно вгнутої функції двох дійсних змінних / М.І.Глебена, Г.Г.Цегелик // Наук. вісник Ужгород. ун-ту. Сер. матем. і інформ. – 2009. – Вип. №18 . – С. 46–50.
3. *Цегелик Г. Г.* Моделювання та оптимізація доступу до інформації файлів баз даних для однопроцесорних та багатопроцесорних систем / Г.Г.Цегелик. –Львів: Видавничий центр ЛНУ імені Івана Франка, – 2010. –192 с.
4. *Цегелик Г. Г.* Математичне моделювання та оптимізація доступу до інформації індексно-послідовних файлів баз даних / Г.Г.Цегелик, А.В.Мельничин // Волин. матем. вісн. Сер. прикл. матем. – 2009. – Вип. №6(15). – С. 179 - 196.

Одержано 10.04.2018