УДК 519.68

Г.Г. Цегелик, А.В. Мельничин
Львівський національний університет імені Івана Франка,
кафедра математичного моделювання соціально-економічних процесів

ЕФЕКТИВНІСТЬ МЕТОДУ R - РІВНЕВОГО БЛОЧНОГО ПОШУКУ ІНФОРМАЦІЇ У ФАЙЛАХ БАЗ ДАНИХ ДЛЯ РІЗНИХ ЗАКОНІВ РОЗПОДІЛУ ЙМОВІРНОСТЕЙ ЗВЕРТАННЯ ДО ЗАПИСІВ

Досліджується ефективність методу r-рівневого блочного пошуку для різних законів розподілу ймовірностей звертання до записів. Для кожного закону розподілу ймовірностей звертання до записів проводиться порівняльний аналіз ефективності методу.

Efficiency of method of r-level block search for different laws of probability distribution of requesting to the records has been investigated. For every law of probability distribution the comparative analysis of efficiency of search method is explored.

Основу сучасних технологій обробки інформації становлять бази даних (БД) і системи керування базами даних (СКБД). Еволюція СКБД відбувається на фоні безпрецедентного росту кількості різноманітних застосувань ЕОМ, а технологія БД, своєю чергою, забезпечує необхідний фундамент такого росту. Основний акцент під час розв'язування різноманітних задач з використанням концепції БД переноситься з процедур обробки інформації на процедури організації збереження та пошуку інформації в БД. Тому продуктивність обчислювальних систем, інформаційним ядром яких є величезні БД, значною мірою визначається ефективністю методів пошуку інформації в файлах БД.

У роботі досліджується ефективність методу *r*-рівневого блочного пошуку як у випадку рівномірного розподілу ймовірностей звертання до записів, так і для різних законів нерівномірного розподілу ймовірностей. За критерій ефективності приймається математичне сподівання кількості порівнянь, необхідних для пошуку запису в файлі. Оскільки в більшості систем обробки інформації типовими є випадки якраз нерівномірного розподілу ймовірностей звертання до записів, то такий підхід дає змогу дослідити ефективність методу для кожного конкретного закону розподілу ймовірностей звертання до записів залежно від зміни параметра r; визначити для кожного закону розподілу ймовірностей значення *r*, для якого математичне сподівання кількості порівнянь, необхідних для пошуку запису, досягає мінімуму; для конкретних *r* дослідити залежність ефективності методу від зміни закону розподілу ймовірностей.

Серед законів нерівномірного розподілу ймовірностей нами розглядаються [1–3]:

– закон Зіпфа

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

де p_i - ймовірність звертання до і - го запису файла,

$H_N = \sum_{k=1}^{N} \frac{1}{k}$ - частинна сума гармонічного ряду,

N - число записів файла;

– узагальнений закон розподілу

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

де c - будь-який параметр $(0 < c < 1)$, $H_N^{(c)} = \sum_{k=1}^{N} \frac{1}{k^c}$ - частинна сума узагальненого гармонічного ряду (при $c = 0{,}8614$ частковим випадком цього закону є розподіл, який наближено задовольняє правило "80–20");

– "бінарний" закон розподілу

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

Треба зауважити, що деякі часткові результати дослідження ефективності цього методу пошуку одержані в [4 – 6].

Розглянемо r-рівневий блочний пошук [4]. Нехай p_i – ймовірність звертання до i-го запису файла. Представимо математичне сподівання кількості порівнянь, необхідних для пошуку запису в файлі, у вигляді суми математичного сподівання кількості порівнянь, необхідних для локалізації блока r-го рівня, математичного сподівання кількості порівнянь, необхідних для локалізації блоку $(r-1)$-го рівня тощо, математичного сподівання кількості порівнянь, необхідних для локалізації блока першого рівня і математичного сподівання кількості порівнянь, необхідних для пошуку запису в локалізованому блоці першого рівня. Тоді математичне сподівання кількості порівнянь, необхідних для пошуку запису в файлі, подамо формулою

$$E = \sum_{i_r=1}^{n_r} \sum_{i_{r-1}=1}^{n_{r-1}} \ldots \sum_{i_1=1}^{n_1} \sum_{i_0=1}^{n_0} (i_r + i_{r-1} + \ldots + i_1 + i_0)\, p_{\varphi(i_0, i_1, \ldots, i_r)},$$

де n_i, $i = 1,\ 2,\ \ldots\ r$, – кількість блоків i-го рівня, n_0 – кількість записів в кожному блоці першого рівня,

$$\varphi(i_0, i_1, \ldots, i_r) = i_0 + \sum_{j=1}^{r} (i_j - 1) \prod_{k=0}^{j-1} n_k .$$

Якщо розподіл імовірностей звертання до записів є рівномірний, то для E одержуємо вираз

$$E = \sum_{k=0}^{r} \frac{1}{2}(n_k + 1).$$

Знайдемо значення параметрів n_k, $k = 1,\ 2,\ \ldots\ r$, за яких E досягає мінімуму.

Покладемо

$$m_i = \prod_{k=0}^{i} n_k, \qquad i = 0,1,\ldots,r.$$

Тоді формулу для E перепишемо у вигляді

$$E = \frac{1}{2} \sum_{k=0}^{r} \left(\frac{m_k}{m_{k-1}} + 1\right),$$

де $m_{-1} = 1,\ m_r = N.$

Для визначення значень параметрів m_i, $i = 0,1,..., r-1$, за яких E досягає мінімуму, одержуємо систему рівнянь

$$m_i^2 = m_{i-1} m_{i+1}, \quad i = 0,1,..., r-1.$$

Із цієї системи маємо

$$m_1 = m_0^2, \quad m_2 = m_0^3, ..., \quad m_{r-1} = m_0^r.$$

Оскільки $m_r = N$, то $m_0^{r+1} = N$. Звідси $m_0 = \sqrt[r+1]{N}$. Тому розв'язком системи рівнянь буде набір величин

$$m_i = N^{\frac{i+1}{r+1}}, \quad i = 0, 1, \ldots, r-1.$$

Отже, параметри n_i $i = 0, 1, \ldots, r$, для яких Е досягає мінімуму, визначаються за формулою

$$n_i = \sqrt[r+1]{N}, \qquad i = 0,1,..., r.$$

При цьому

$$E_{\min} = \frac{1}{2}(r+1)(\sqrt[r+1]{N} + 1).$$

r-рівневий блочний пошук називатимемо r-рівневим блочним пошуком з оптимальним розміром блоків на всіх рівнях, якщо на i-му рівні ($i = 1, 2, \ldots, r$) кожний блок розбивається на $n_i = \sqrt[r+1]{N}$ підблоків по $s_i = \sqrt[r+1]{N}$ записів у кожному.

Знайдемо тепер r, для якого E_{min} набуватиме найменшого значення. Неважко переконатися, що E_{min} досягає мінімуму при

$$r = r_{on} = \frac{\ln N}{\ln t_{on}} - 1,$$

де $t_{on} = 3.59$ – додатний корінь рівняння

$$1 + \frac{1}{t} = \ln t.$$

Нехай розподіл ймовірностей звертання до записів задовольняє закон Зіпфа.

Тоді

$$p_{\varphi(i_0, i_1, \ldots, i_r)} = \frac{1}{\varphi(i_0, i_1, \ldots, i_r) H_N}$$

і для E аналогічно як в [2] отримуємо вираз

$$E = \frac{1}{H_N}((n_r + 1)H_N - S_{m_{r-1}}(l_r) + n_0 S_{m_0}(l_1) - N(H_N - 1) +$$

$$+ \sum_{k=1}^{r-1}(n_k S_{m_k}(l_{k+1}) - S_{m_{k-1}}(l_k) + H_N)),$$

де

$$m_i = \prod_{k=0}^{i} n_k, \quad l_i = \prod_{k=i}^{r} n_k, \quad i = 0,1,..., r,$$

$$S_{m_i}(l_{i+1}) = \sum_{k=1}^{l_{i+1}} H_{km_i}, \qquad i = 0,1,...,r-1.$$

Якщо використати апроксимацію $S_{m_{i-1}}(l_i)$ виразом [2]

$$\bar{S}_{m_{i-1}}(l_i) = l_i(H_N - 1) + \frac{1}{2}\ln l_i + C_1, \quad i = 1,2,...,r,$$

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

$$E = \frac{1}{H_N}(H_N + l_r - \frac{1}{2}\ln l_r - C_1 + \frac{N}{l_1}(\frac{1}{2}\ln l_1 + C_1) + \sum_{k=1}^{r-1}(H_N + (\frac{l_k}{l_{k+1}} - 1)(\frac{1}{2}\ln l_{k+1} + C_1) - \frac{1}{2}\ln\frac{l_k}{l_{k+1}})),$$

де $l_0 = N$, $l_{r+1} = 1$, $m_{-1} = 1$, $m_r = N$.

Для знаходження значень параметрів l_i, $i = 1, 2, ..., r$, для яких E досягає мінімуму, одержимо таку систему нелінійних рівнянь

$$\begin{cases} l_{r-1} = \dfrac{(2l_r - 1)l_r}{\ln l_r + 2C_1 - 1}, \\ l_{k-1} = \dfrac{(\frac{l_k}{l_{k+1}}(\ln l_{k+1} + 2C_1) - 1)l_k}{\ln l_k + 2C_1 - 1} \quad (k = 2,3,...,r-1), \\ \dfrac{l_1}{l_2}(\ln l_2 + 2C_1) - 1 = \dfrac{N}{l_1}(\ln l_1 + 2C_1 - 1). \end{cases}$$

Припустимо, що розподіл ймовірностей звертання до записів задовольняє узагальнений закон. Тоді

$$p_{\varphi(i_0,i_1,...,i_r)} = \frac{1}{(\varphi(i_0,i_1,...,i_r))^c H_N^{(c)}}$$

і для E аналогічно як в [2]одержуємо вираз

$$E = \frac{1}{H_N^{(c)}}((n_r + 1)H_N^{(c)} - S_{m_{r-1}}^{(c)}(l_r) + n_0 S_{m_0}^{(c)}(l_1) + H_N^{(c-1)} - NH_N^{(c)} +$$

$$+ \sum_{k=1}^{r-1}(n_k S_{m_k}^{(c)}(l_{k+1}) - S_{m_{k-1}}^{(c)}(l_k) + H_N^{(c)})),$$

де

$$m_i = \prod_{k=0}^{i} n_k, \quad l_i = \prod_{k=i}^{r} n_k, \quad i = 0,1,...,r,$$

$$S_{m_i}^{(c)}(l_{i+1}) = \sum_{k=1}^{l_{i+1}} H_{km_i}^{(c)}, \qquad i = 0,1,...,r-1.$$

Якщо використати апроксимацію $S_{m_{i-1}}^{(c)}(l_i)$ виразом [2]

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

де

$$\alpha^{(c)}(l_i) = H_{l_i}^{(c-1)} - \frac{1}{2-c}l_i^{2-c}, \quad i = 1,2,...,r,$$

то з достатньо високою точністю можемо прийняти

$$E = \frac{1}{H_N^{(c)}}\left(H_N^{(c)} + \frac{N^{1-c}}{1-c}\left(\frac{1-c}{2-c}l_r - \frac{\alpha^{(c)}(l_r)}{l_r^{1-c}}\right) + H_N^{(c-1)} - \frac{N^{2-c}}{1-c}\left(\frac{1-c}{2-c} - \frac{\alpha^{(c)}(l_1)}{l_1^{2-c}}\right) + \right.$$
$$\left. + \sum_{k=1}^{r-1}\left(H_N^{(c)} + \frac{N^{1-c}}{1-c}\left(\frac{l_k\alpha^{(c)}(l_{k+1})}{l_{k+1}^{2-c}} - \frac{\alpha^{(c)}(l_k)}{l_k^{1-c}}\right)\right)\right).$$

Для знаходження наближених значень параметрів l_i, за яких E досягає мінімуму, одержуємо таку систему нелінійних рівнянь (при цьому похідні від $\alpha^{(c)}(l_i)$ за l_i замінені скінченними різницями $\alpha^{(c)}(l_i+1) - \alpha^{(c)}(l_i)$):

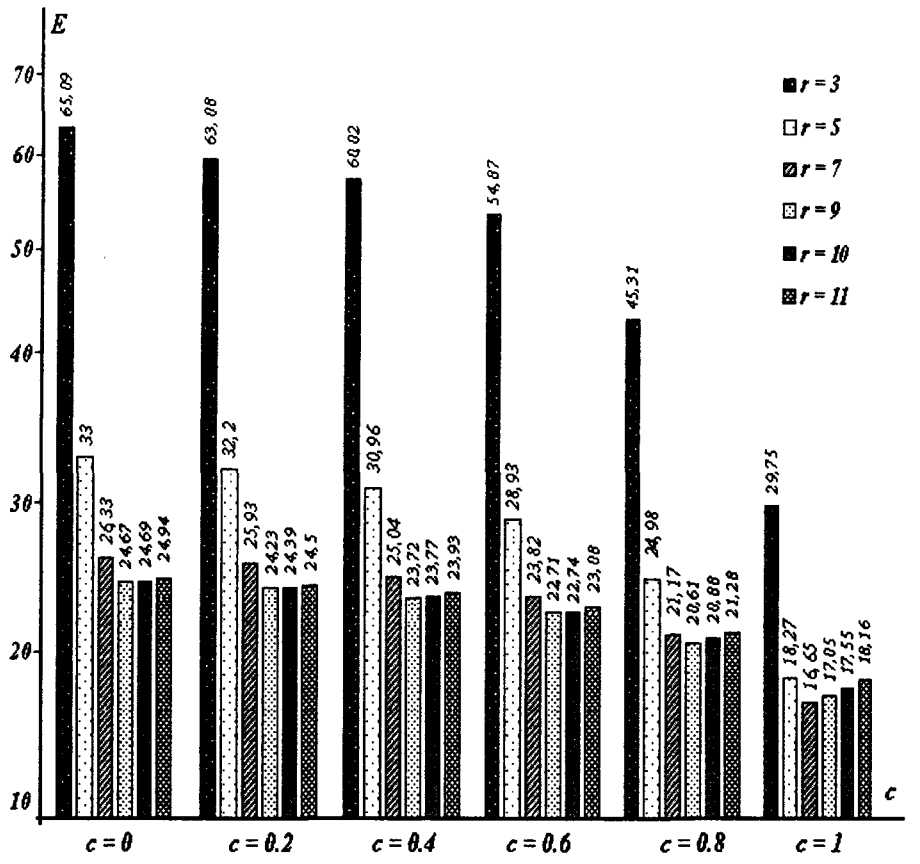
$$\begin{cases} l_{r-1} = \dfrac{(\alpha^{(c)}(l_r+1) - \alpha^{(c)}(l_r))l_r^2 - (1-c)\alpha^{(c)}(l_r)l_r - \dfrac{1-c}{2-c}l_r^{3-c}}{(\alpha^{(c)}(l_r+1) - \alpha^{(c)}(l_r))l_r - (2-c)\alpha^{(c)}(l_r)}, \\ l_{k-1} = \dfrac{(\alpha^{(c)}(l_k+1) - \alpha^{(c)}(l_k))l_k^2 - (1-c)\alpha^{(c)}(l_k)l_k - \dfrac{l_k^{3-c}}{l_{k+1}^{2-c}}\alpha^{(c)}(l_{k+1})}{(\alpha^{(c)}(l_k+1) - \alpha^{(c)}(l_k))l_k - (2-c)\alpha^{(c)}(l_k)} \quad (k = 2,3,...,r-1), \\ (\alpha^{(c)}(l_1+1) - \alpha^{(c)}(l_1))l_1^2 - (1-c)\alpha^{(c)}(l_1)l_1 - \\ - N((\alpha^{(c)}(l_1+1) - \alpha^{(c)}(l_1))l_1 - (2-c)\alpha^{(c)}(l_1)) = \dfrac{l_1^{3-c}}{l_2^{2-c}}\alpha^{(c)}(l_2). \end{cases}$$

Оптимальні значення r (з точністю до 0,1) для деяких N та для різних законів розподілу ймовірностей звертання до записів приведені у таблиці.

Зауважимо, що у випадку рівномірного закону розподілу ймовірностей оптимальне значення r обчислене на основі виведеного для r_{on} виразу, а у разі закону Зіпфа та узагальненого закону розподілу (для різних c) інтервали для r_{on} знайдено з використанням числового експерименту.

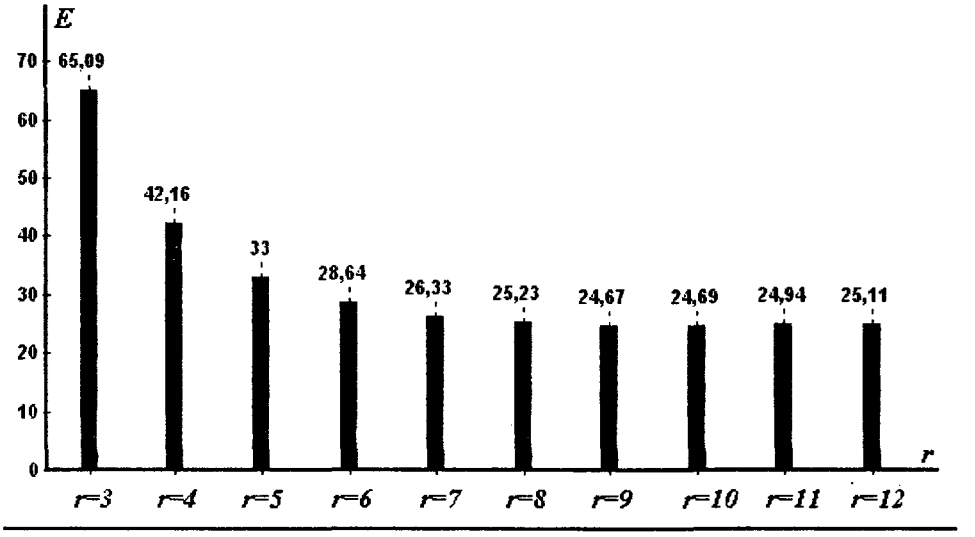
N	10^4	10^5	10^6
c=0	6.2	8.0	9.8
c=0.2	(5.7 ; 6.8)	(7.9 ; 8.5)	(9.9 ; 10.5)
c=0.4	(5.4 ; 6.5)	(7.2 ; 8.1)	(9.7 ; 10.1)
c=0.6	(5.3 ; 6.1)	(6.9 ; 7.9)	(9.2 ; 9.9)
c=0.8	(4.9 ; 5.7)	(6.2 ; 7.3)	(8,8 ; 9.3)
c=1	(4.6 ; 5.3)	(5.8 ; 6.1)	(8.1 ; 9.0)

На гістограмі 1 показано залежність математичного сподівання кількості порівнянь, необхідних для пошуку запису в файлі, від зміни закону розподілу ймовірностей звертання до записів для різних значень параметра r при $N = 10^6$.



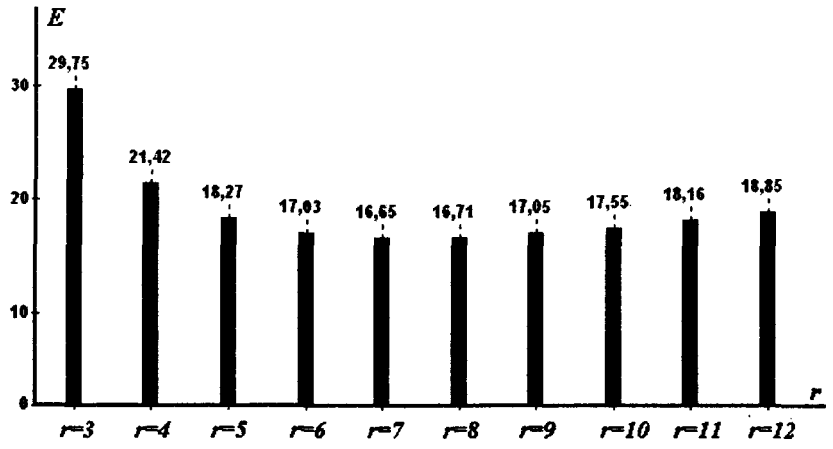
Гістограма 1

На гістограмі 2 наведено залежність математичного сподівання кількості порівнянь, необхідних для пошуку запису в файлі, від зміни параметра r за $N = 10^6$ у випадку рівномірного закону розподілу ймовірностей звертання до записів.



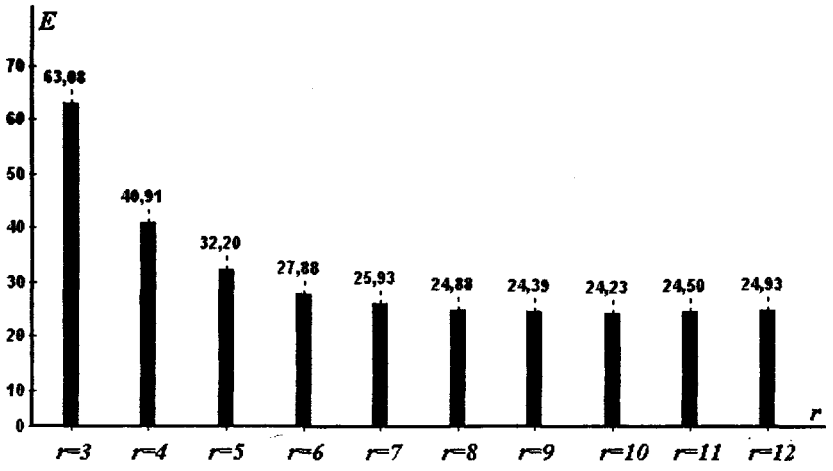
Гістограма 2

На гістограмі 3 наведено залежність математичного сподівання кількості порівнянь, необхідних для пошуку запису в файлі, від зміни параметра r за $N = 10^6$ у випадку розподілу ймовірностей звертання до записів за законом Зіпфа.



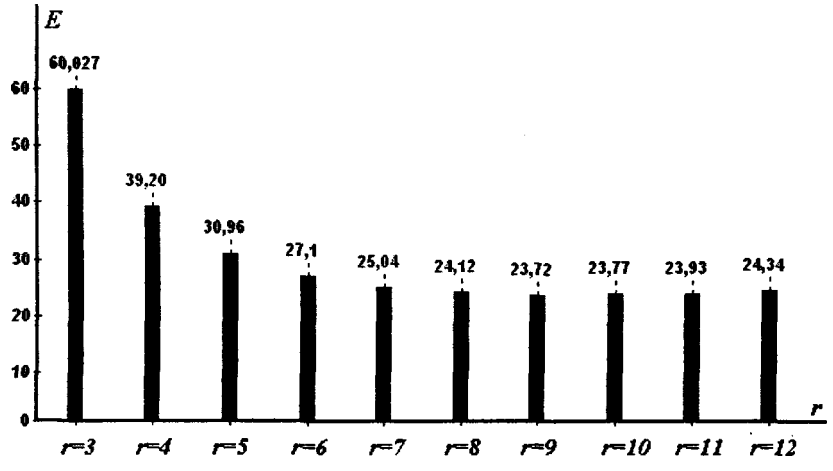
Гістограма 3

Ефективність методу r-рівневого блочного пошуку для різних r у випадку узагальненого закону розподілу ймовірностей звертання до записів для $c = 0.2$ і $N = 10^6$ показано на гістограмі 4.



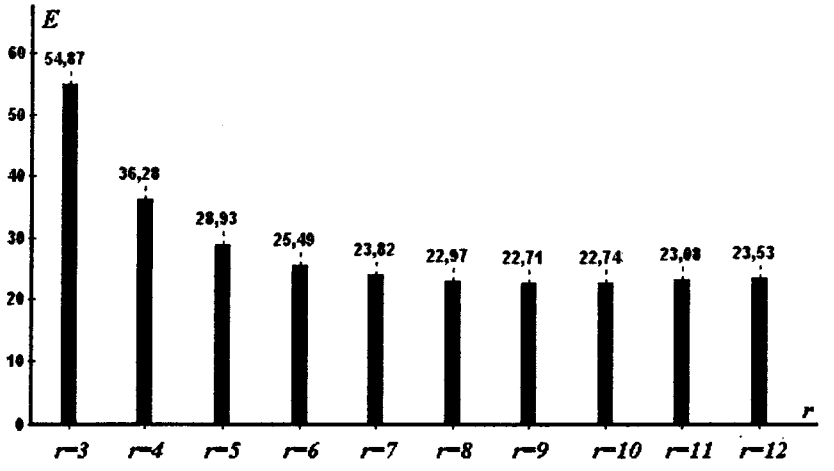
Гістограма 4

На гістограмі 5 наведено залежність математичного сподівання кількості порівнянь, необхідних для пошуку запису в файлі, від зміни параметра r у випадку узагальненого закону розподілу ймовірностей звертання до записів для $c = 0.4$ і $N = 10^6$.



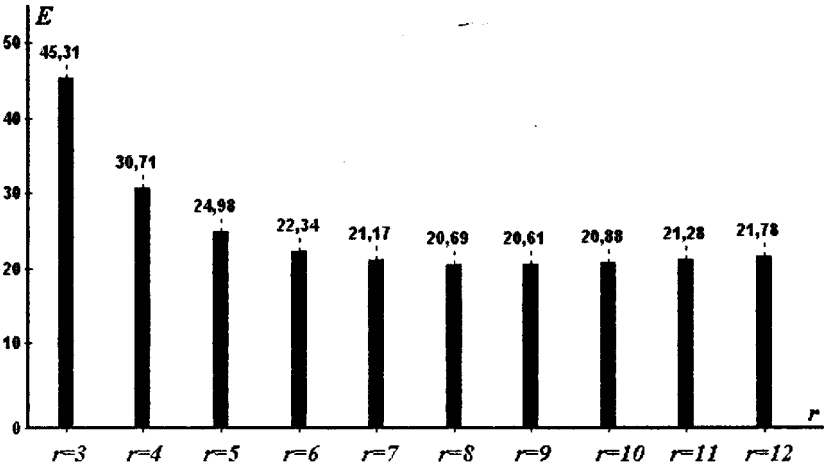
Гістограма 5

На гістограмі 6 наведено залежність математичного сподівання кількості порівнянь, необхідних для пошуку запису в файлі, від зміни параметра r у випадку узагальненого закону розподілу ймовірностей звертання до записів для $c = 0.6$ і $N = 10^6$.



Гістограма 6

Ефективність методу r-рівневого блочного пошуку для різних r у випадку узагальненого закону розподілу ймовірностей звертання до записів для $c = 0.8$ і $N = 10^6$ показано на гістограмі 7.



Гістограма 7

Якщо розглядати випадок розподілу ймовірностей звертання до записів за „бінарним" законом, то неважко переконатися, що математичне сподівання кількості порівнянь, необхідних для пошуку запису в файлі, зростає за зростання r. Тому в цьому випадку r-рівневий блочний пошук буде оптимальним за $r = 1$. Тоді аналогічно як в [2] з точністю до нескінченно малої величини

$$E = \frac{2^m}{2^m - 1}\left(3 - \frac{m+2}{2^m}\right).$$

Звідси випливає, що E не залежить від N і досягає мінімуму при m, що є додатним коренем рівняння

$$2^m\left((m-1)\ln 2 - 1\right) + 1 = 0.$$

Оскільки додатним коренем цього рівняння є $m = m_o = 2{,}1$ (з точністю до 0.1), то з точністю до 0.01 $E_{min} = 2{,}84$.

Одержані результати за дослідження ефективності методу r - рівневого блочного пошуку записів у файлах для різних законів розподілу ймовірностей звертання до записів можуть бути використані для удосконалення механізмів доступу до інформації баз даних.

1. Кнут Д. Искусство программирования для ЭВМ. Т.3. Сортировка и поиск. – М.: Издательский дом "Вильямс", 2000. – 832 с. 2.Цегелик Г.Г. Организация и поиск информации в базах данных. – Львов: Вища школа, 1967. – 176 с. 3.Цегелик Г.Г. Системы распределенных баз данных. – Львов: Свит, 1990. – 168 с. 4. Мартин Дж. Организация баз данных в вычислительных системах. – М.: Мир, 1980. – 644 с. 5.Філяк М.І., Цегелик Г.Г. Метод r - рівневого блочного пошуку записів у впорядкованих файлах і його ефективність // Вісн. Львів. ун-ту. Сер. "Прикл. матем. та інформ.". – 2000. – Вип. 3. – С. 169 – 173. 5. Цегелик Г.Г., Філяк М.І. Про ефективність методу r - рівневого блочного пошуку записів у впорядкованих файлах // Вісн. Львів. ун-ту. Сер. "Прикл. матем. та інформ.". – 2002. – Вип. 5. – С. 174 – 177.

УДК 621.3.019.3:519.2

С. Ю. Юриш
Національний університет "Львівська політехніка",
кафедра інформаційних систем та мереж

ТОЧНИЙ МЕТОД ДОДАВАННЯ СКЛАДОВИХ ПОХИБОК НА ОСНОВІ ВИКОРИСТАННЯ ХАРАКТЕРИСТИЧНИХ ФУНКЦІЙ І ЛІНІЙНО-ЛАМАНОЇ АПРОКСИМАЦІЇ ЗАКОНІВ РОЗПОДІЛУ

Описано метод побудови закону розподілу густини імовірності і визначення числових характеристик результуючої похибки n складових з наперед заданою точністю. Метод базується на використанні характеристичних функцій А.М. Ляпунова. На відміну від раніше запропонованого комп'ютерного методу побудови густини ймовірності у новому методі для представлення закону розподілу замість східчастої апроксимації використовується лінійно-ламана апроксимація з автоматичною зміною кроку дискретизації залежно від заданої точності. Це дало змогу отримати оптимальне співвідношення швидкодія/точність під час реалізації алгоритму цього методу на ПЕОМ.

The method for construction of the distribution law of probability density function and numerical characteristics determination for the resulting error of n components with beforehand given accuracy are described in this article. The method is based on Lyapunov's characteristic functions. For distribution law presentation the linear-broken approximation instead of step approximation is used. The method has an automatic change of discretization step depends on the given accuracy. It has allowed to achieve an optimum speed/accuracy ratio at its PC-based realization.

Вступ

Задача розрахункового додавання складових похибок являє собою одну з основних задач метрології під час створення засобів вимірювань, а також оцінювання похибок результатів самих вимірів [1]. Загальний підхід до визначення густини розподілу суми похибок вимірів полягає в зведенні задачі до обчислення розподілу суми незалежних випадкових величин та застосуванні для розв'язання останньої методів теорії ймовірності. Такий метод додавання з урахуванням усіх