ISSN 2078–5097. Вісн. Львів. ун-ту. Сер. прикл. матем. та інф. 2013. Вип. 20. С. 95–99
Visnyk of the Lviv University. Series Appl. Math. and Informatics. Issue 20. P. 95–99

УДК 004.451.53: 004.625.5

ПОРІВНЯЛЬНИЙ АНАЛІЗ ОПТИМАЛЬНИХ МОДЕЛЕЙ ІНДЕКСУ В ІНДЕКСНИХ МЕТОДАХ ОРГАНІЗАЦІЇ ФАЙЛІВ БАЗ ДАНИХ

А. Мельничин, Г. Цегелик

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

Визначено параметри, за яких математичне сподівання кількості порівнянь, потрібних для пошуку елемента в індексі, організованого у вигляді повністю збалансованого індексного дерева, досягає мінімуму для заданої кількості елементів індексу. Проведено порівняльний аналіз оптимальних моделей індексу для різних законів розподілу ймовірностей звертання до його елементів.

Ключові слова: індексні методи організації файлів баз даних.

1. ВСТУП

Найпоширеніші методи організації файлів баз даних – індексні методи [1]. Найважливіша задача, яка виникає під час реалізації таких методів, це задача ефективної організації індексу. Здебільшого індекс організовується у вигляді багаторівневого збалансованого індексного дерева. Оптимальні моделі такого дерева досліджено в [2] у випадку рівномірного розподілу ймовірностей звертання до елементів індексу за використання різних методів пошуку у його вузлах.

Якщо розподіли ймовірностей звертання до елементів вузлів дерева нерівномірні (“бінарний”, закон Зіпфа, узагальнений розподіл), то в [3] знайдено явний вигляд математичного сподівання кількості порівнянь, потрібних для пошуку елемента в індексі. При цьому у випадку “бінарного” розподілу ймовірностей показано, що математичне сподівання досягає мінімуму тільки тоді, коли індексне дерево є однорівневим, у разі закону Зіпфа виведено рівняння для визначення параметрів, за яких математичне сподівання досягає мінімуму.

2. ФОРМУЛЮВАННЯ ЗАДАЧІ

Припустимо, що індекс, який містить N елементів, організований у вигляді повністю збалансованого індексного дерева. Нехай l – кількість елементів у кожному вузлі дерева, r – кількість рівнів вузлів у дереві, p_i – ймовірність звертання до i-го елемента індексу. Тоді за використання методу послідовного перегляду у вузлах дерева математичне сподівання кількості порівнянь, потрібних для пошуку елемента в індексі, виражається формулою [3]

$$E = \sum_{i_r=1}^{l} \sum_{i_{r-1}=1}^{l} \cdots \sum_{i_1=1}^{l} (i_1 + i_2 + \ldots + i_r) p_{\varphi(i_1, i_2, \ldots, i_r)},$$

де $\varphi(i_1, i_2, \ldots, i_r) = i_1 + \sum_{j=2}^{r} (i_j - 1)\, l^{j-1}$. В [3] виведено явний вигляд E для конкретних законів розподілу ймовірностей звертання до елементів індексу.

Знайдемо значення параметрів l і r, за яких E досягає мінімуму, і проведемо порівняльний аналіз оптимального значення E для різних законів розподілу ймовірностей звертання до елементів індексу.

3. РОЗВ'ЯЗАННЯ ЗАДАЧІ

1. Якщо розподіл ймовірностей звертання до елементів індексу рівномірний, то [3]

$$E = \frac{(l+1)\ln N}{2\ln l}$$

і для визначення параметра l, за якого E досягає мінімуму, отримаємо рівняння

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

Коренем цього рівняння з точністю до 0.1 є $l = l_0 = 3{,}6$.

Оптимальні значення математичного сподівання для різних N з точністю до 0.001 наведено в табл. 1.

Таблиця 1

Оптимальні значення математичного сподівання

N	10^3	10^4	10^5	10^6	10^7	10^8	10^9
E_{on}	12,403	16,538	20,672	24,807	28,941	33,076	37,210

2. Нехай розподіл імовірностей звертання до елементів індексу є "бінарним" [4-5]. Тоді [3]

$$E = \left(r + 1 - (l-1)\sum_{i=1}^{r-1} \frac{1}{2^{l^i} - 1} \right)\left(1 - 2^{-N}\right) + (r-1)l2^{-N}.$$

У [3] показано, що E досягає мінімуму для $r = 1$. У цьому випадку $E = 2(1 - 2^{-N})$ і з точністю до нескінченно малої $E_{on} \approx 2$.

3. Припустимо, що розподіл імовірностей звертання до елементів індексу задовольняє закон Зіпфа [4-5]. Тоді з достатньо високою точністю [3]

$$E = \left(1 + (l-1)h_1\right)\frac{\ln N}{\ln l} + lh_2 + h_1 - 1,$$

де $h_1 = \frac{1}{H_N}\left(\frac{1}{4}\ln N + C_1\right)$, $h_2 = \frac{1}{H_N} - h_1$, $C_1 = \frac{1}{2}\ln 2\pi$.

Для знаходження наближеного значення l, за якого E досягає мінімуму, одержимо рівняння

$$(\ln l - 1)h_1 + \frac{h_2}{\ln N}\ln^2 l = \frac{1}{l}(1 - h_1).$$

Корені цього рівняння і відповідні значення математичного сподівання для різних значень N наведено у табл. 2.

Таблиця 2

Оптимальні значення математичного сподівання

N	10^3	10^4	10^5	10^6	10^7	10^8	10^9
l	4,952	4,967	4,972	4,975	4,976	4,976	4,976
E_{on}	8,614	11,451	14,297	17,148	20,002	22,858	25,714

4. Нехай розподіл імовірностей звертання до елементів індексу задовольняє узагальнений закон розподілу [5]. Тоді в [3] для наближеного обчислення E виведена формула

$$E = \frac{1}{H_N^{(c)}}\left((r-1)H_N^{(c)} + H_N^{(c-1)} - \left(\frac{N-\sqrt[r]{N}}{2-c} - \frac{\sqrt[r]{N}-1}{1-c}\delta(r)\right)N^{1-c}\right),$$

де

$$0 < c < 1,\ H_N^{(c)} = \sum_{i=1}^{N} i^{-c},\ \delta(r) = \sum_{i=1}^{r-1} \frac{\alpha(N^{1-i/r})}{(N^{1-i/r})^{1-c}},$$

$\alpha(x)$ – деяка повільно зростаюча функція.

У табл. 3-5 подано значення функції E для різних значень параметра c, деяких конкретних значень r та N.

Таблиця 3

Значення функції E для $N = 10^3$

N	***c***	*r*					
		2	**3**	**4**	**5**	**6**	**7**
10^3	0.2	30.652	15.735	12.733	12.035	12.118	12.550
	0.4	28.128	14.738	12.063	11.493	11.642	12.113
	0.6	24.842	13.406	11.166	10.770	11.009	11.534
	0.8	20.705	11.685	10.004	9.834	10.192	10.790

Таблиця 4

Значення функції E для $N = 10^6$

N	***c***	*r*					
		8	**9**	**10**	**11**	**12**	**13**
10^6	0.2	25.959	24.912	24.471	24.406	24.590	24.946
	0.4	25.200	24.239	23.857	23.853	24.051	24.432
	0.6	23.935	23.119	22.836	22.887	23.159	23.583
	0.8	21.472	20.937	20.851	21.046	21.128	21.938

Таблиця 5

Значення функції E для $N = 10^9$

N	***c***	*r*					
		14	**15**	**16**	**17**	**18**	**19**
10^9	0.2	37.299	36.924	36.800	36.865	37.077	37.404
	0.4	36.644	36.307	36.214	36.304	36.536	36.880
	0.6	35.516	35.244	35.204	35.338	35.606	35.982
	0.8	32.816	32.403	32.791	33.032	33.391	33.843

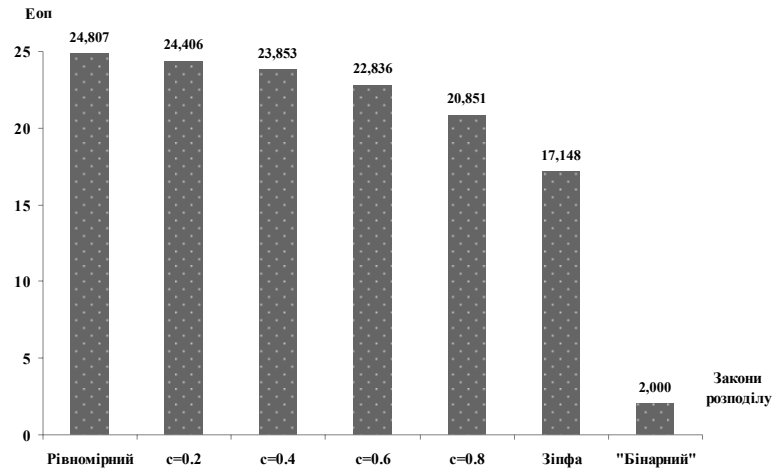
На підставі даних табл. 3-5 можна визначити оптимальні значення математичного сподівання E_{on} для деяких c та N. Значення E_{on} подано в табл. 6.

Таблиця 6

Оптимальні значення математичного сподівання

N	*c*			
	0.2	**0.4**	**0.6**	**0.8**
10^3	12.035	11.493	10.770	9.834
10^6	24.406	23.853	22.836	20.851
10^9	36.800	36.214	35.204	32.403

Маючи оптимальні значення математичного сподівання для різних законів розподілу ймовірностей звертання до елементів індексу, можемо їх порівняти для деяких N. Для $N = 10^6$ ці порівняння зображені на рис.



Порівняння E_{on} для $N = 10^6$ і різних законів розподілу ймовірностей звертання до елементів індексу

4. ВИСНОВКИ

Розглянуто організацію індексу у вигляді повністю збалансованого індексного дерева. Визначено параметри, за яких математичне сподівання кількості порівнянь, потрібних для пошуку елемента в індексі, досягає мінімуму, для заданої кількості елементів індексу. Проведено порівняльний аналіз оптимальних моделей індексу для різних законів розподілу ймовірностей звертання до його елементів. За зміни закону розподілу ймовірностей від рівномірного до закону Зіпфа, оптимальні значення математичного сподівання зменшуються, але не суттєво. Для "бінарного" розподілу ймовірностей воно значно відрізняється від інших випадків.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. *Мартин Дж.* Организация баз данных в вычислительных системах / Дж. Мартин. – М.: Мир, 1980. – 664 с.
2. *Цегелик Г.Г.* Организация и поиск информации в базах данных / Г.Г. Цегелик. – Львов: Вища школа, 1987. – 176 с.

3. *Цегелик Г.Г.* Оптимальные модели индекса в индексных методах организации файлов баз данных / Г.Г. Цегелик // Науч. сб. “Модели и системы обработки информации”. – 1990. – Вып. 9. – С. 28–32.
4. *Кнут Д.* Искусство программирования для ЭВМ. Т. 3: Сортировка и поиск. / Д. Кнут. – М.: Изд. дом “Вильямс”, 2000. – 832 с.
5. *Цегелик Г.Г.* Моделювання та оптимізація доступу до інформації файлів баз даних для однопроцесорних і багатопроцесорних систем / Г.Г. Цегелик. – Львів: Видавничий центр ЛНУ ім. Івана Франка, 2010. – 192 с.

Стаття: надійшла до редколегії 26.09.2012
доопрацьована 21.11.2012
прийнята до друку 05.12.2012

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ОПТИМАЛЬНЫХ МОДЕЛЕЙ ИНДЕКСА В ИНДЕКСНЫХ МЕТОДАХ ОРГАНИЗАЦИИ ФАЙЛОВ БАЗ ДАННЫХ

А. Мельничин, Г. Цегелик

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

Определены параметры, при которых математическое ожидание числа сравнений, необходимых для поиска элемента в индексе, организованном в виде полностью сбалансированного индексного дерева, достигает минимума, для заданного количества элементов индекса. Проведен сравнительный анализ оптимальных моделей индекса для различных законов распределения вероятностей обращение к его элементам.

Ключевые слова: индексные методы организации файлов баз данных.

COMPARATIVE ANALYSIS OPTIMAL MODELS INDEX METHODS ORGANIZATION OF DATABASE FILES

A. Melnychyn, G. Tsegelyk

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

Parameters for which the mathematical expectation number of comparisons required to search for an item in index, organized as a fully balanced index-tree, reaches a minimum, for a given amount element of the index are defined. A comparative analysis of the optimal models of the index for different probability distributions of appeals to the elements was carried.

Key words: index methods organization database files.