К.т.н. Мельничин А.В., д.ф.-м.н. Цегелик Г.Г.

Львівський національний університет імені Івана Франка, Україна

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

Найбільш поширеними методами організації файлів баз даних є індексні методи. Однією з найважливіших задач, що виникають при реалізації таких методів, є задача ефективної організації індексу. У більшості випадків індекс організовується у вигляді багаторівневого збалансованого індексного дерева. Оптимальні моделі такого дерева досліджено в [1] для рівномірного розподілу ймовірностей звертання до елементів індексу у випадку використання різних методів пошуку у його вузлах.

Якщо розподіли ймовірностей звертання до елементів вузлів дерева є нерівномірні (“бінарний”, закон Зіпфа, узагальнений розподіл), то в [2] знайдено явний вигляд математичного сподівання кількості порівнянь, необхідних 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; в разі закону Зіпфа виведено рівняння для визначення параметрів, за яких математичне сподівання досягає мінімуму.

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, необхідних для пошуку елемента в індексі, виражається формулою

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

де $\varphi(i_1, i_2, \ldots, i_r) = i_1 + \sum_{j=2}^{r} (i_j - 1) l^{j-1}$.

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

1. Якщо розподіл ймовірностей звертання до елементів індексу є рівномірним, то

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

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

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

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

2. Нехай розподіл ймовірностей звертання до елементів індексу є "бінарним". Тоді

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

і E досягає мінімуму для $r = 1$. У цьому випадку $E = 2(1 - 2^{-N})$ і з точністю до нескінченно малої $E_{\tilde{n}\ddot{i}} = 2$.

3. Припустимо, що розподіл ймовірностей звертання до елементів індексу задовольняє закон Зіпфа. Тоді з достатньо високою точністю

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

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, маємо рівняння

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

4. Нехай розподіл ймовірностей звертання до елементів індексу задовольняє узагальнений закон розподілу. Тоді для наближеного обчислення E використовується 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.

Обчисливши оптимальні значення математичного сподівання для деяких N і різних законів розподілу ймовірностей звертання до елементів індексу, можемо їх порівняти. Так для $N = 10^6$ ці порівняння наведені на рис. 1.

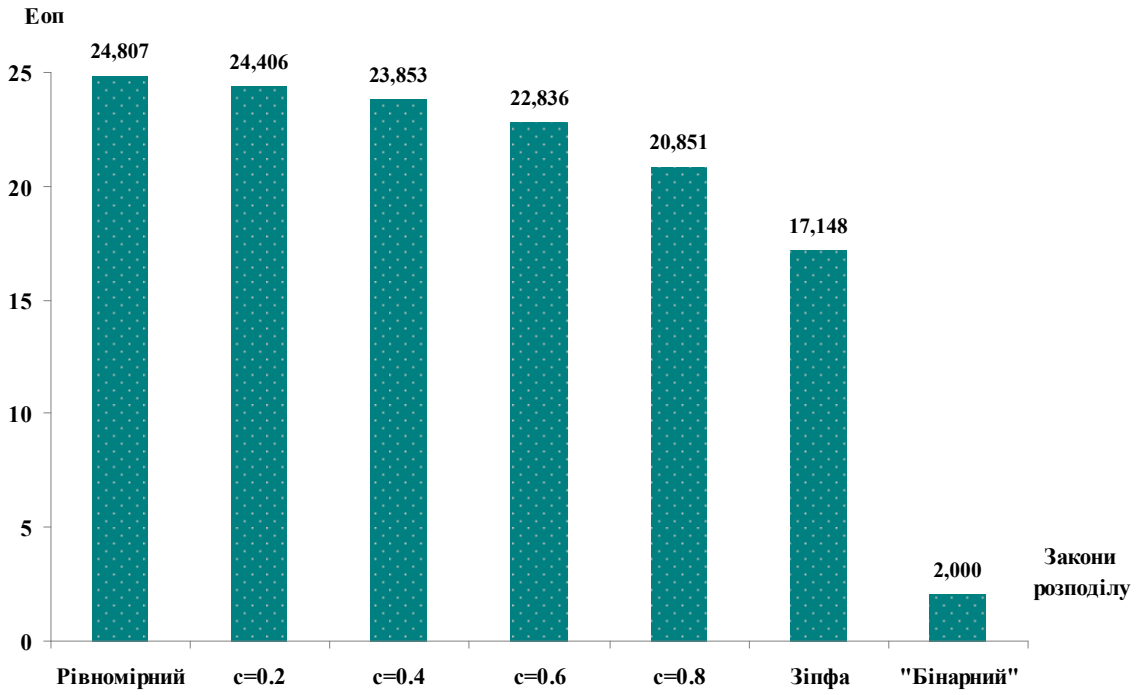

Рис. 1. Порівняння E_{on} для $N = 10^6$ і різних законів розподілу ймовірностей звертання до елементів індексу

Як бачимо, з рис. 1, зі зміною 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 використаних джерел:

1. Цегелик Г.Г. Организация и поиск информации в базах данных / Г.Г. Цегелик. – Львов: Высш. школа, 1987. – 176 с.
2. 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.