Дослідження ефективності методів паралельного пошуку інформації в файлах баз даних

Володимир Лісовець

Кафедра математичного моделювання соціально-економічних процесів, Львівський національний університет ім. І. Франка, УКРАЇНА, м.Львів, вул. Університетська 1, E-mail: kafmmsep@franko.lviv.ua

Abstract **– *The m-parallel method of sequential field searching and two variants of m-parallel block field searching method are considered. We research the effectiveness of these methods for different probability distribution law of field access. The mathematical expectation of number of parallel comparison necessary for field searching in files is taken as a criterion of effectiveness. The effectiveness of the methods is compared and analyzed. Optimal strategy of field searching in sequenced files stored in external memory of multiprocessing system is made. In this case the mathematical expectation of total time needed for field searching in files is taken as a criterion of effectiveness.***

Ключові слова – multiprocessing system, mathematical modeling, parallel searching, database.

I. Методи паралельного пошуку та їх ефективність

Розглядаються наступні методи паралельного пошуку записів у файлах БД для багатопроцесорних ЕОМ [1-4]:

- метод *m*-паралельного послідовного перегляду;
- перший варіант методу *m*-паралельного блочного пошуку;
- другий варіант методу *m*-паралельного блочного пошуку.

Проводиться аналіз ефективності цих методів для різних законів розподілу ймовірностей звертання до записів (рівномірного, «бінарного», Зіпфа й узагальненого, частковим випадком якого є розподіл, що наближено задовольняє правило «80-20» [5-8]). За критерій ефективності приймається математичне сподівання кількості паралельних порівнянь, необхідних для пошуку запису у файлі.

Розглянемо **метод *m*-паралельного послідовного перегляду**.

Припустимо, що до складу багатопроцесорної ЕОМ входять m процесорів, які працюють паралельно та мають спільне поле пам'яті. Пронумеруємо процесори натуральними числами від 1 до m. Суть методу m-паралельного послідовного перегляду полягає в такому. Розіб'ємо всі записи файлу умовно на блоки по m записів у кожному. Нехай $N = n \cdot m$ – кількість записів у файлі, де n – кількість блоків. Тоді при використанні m-паралельного послідовного перегляду процес пошуку запису буде складатися з низки кроків. На першому кроці i-ий процесор переглядає значення ключа i-го запису. При цьому процес перегляду може бути успішним або неуспішним. Для визначення «успішності» всі процесори повинні обмінятися інформацією. У разі успішного перегляду процес пошуку завершується. Якщо перегляд неуспішний, то на другому кроці i-ий процесор переглядає значення ключа $(m + i)$-го запису і т. д. На $(k + 1)$-му кроці (у випадку неуспішного перегляду на k-му кроці) i-ий процесор переглядає значення ключа $(km + i)$-го запису. В наслідок виконання не більше ніж n кроків шуканий запис буде знайдено, якщо він міститься у файлі. Якщо p_i – ймовірність звертання до i-го запису файла, то математичне сподівання E кількості паралельних порівнянь, необхідних для пошуку запису у файлі, виражається формулою

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

Нами знайдений явний вираз метематичного сподівання кількості паралельних порівнянь у випадку різних законів розподілу ймовірностей звертання до записів.

У випадку, **першого варіанту методу *m*-паралельного блочного пошуку** вважаємо, що записи впорядкованого файла розбиті на n блоків по sm записів у кожному і пошук запису у файлі здійснюємо таким чином. Спочатку шукаємо блок, який містить шуканий запис, шляхом перегляду m останніх записів блоків. Після цього пошук продовжуємо у локалізованому блоці за допомогою методу m-паралельного послідовного перегляду. Математичне сподівання E кількості паралельних порівнянь, необхідних для пошуку запису у файлі, запишемо у вигляді суми математичного сподівання кількості паралельних порівнянь, необхідних для локалізації блоку, який містить шуканий запис, і математичне сподівання кількості паралельних порівнянь, необхідних для пошуку запису в локалізованому блоці. Тоді

$$E = \sum_{k=1}^{n} \sum_{i=1}^{s} \sum_{j=1}^{m} (k+i) p_{(k-1)ms+(i-1)m+j} \ . \qquad (2)$$

Нами знайдений явний вираз метематичного сподівання кількості паралельних порівнянь у випадку різних законів розподілу ймовірностей звертання до записів. Та обчислені оптимальні розміри блоків за яких математичне сподівання досягає мінімуму.

У разі **другого варіанту методу *m*-паралельного блочного пошуку** приймаємо, що записи файла розбиті на nm блоків по sm записів у кожному, тоді кількість записів у файлі буде $N = snm^2$. Пошук запису у файлі здійснюється наступним чином. Спочатку шукається блок, який містить шуканий запис, використовуючи метод m-паралельного послідовного перегляду серед останніх елементів блоків. Після цього пошук продовжується в локалізованому блоці також за допомогою методу m-паралельного послідовного перегляду. Математичне сподівання E кількості паралельних порівнянь, необхідних для пошуку запису у файлі, представимо у вигляді суми математичного сподівання кількості паралельних порівнянь, необхідних для локалізації блоку записів, і математичного сподівання кількості паралельних

порівнянь, необхідних для пошуку запису в локалізованому блоці:

$$E = \sum_{k=1}^{n}\sum_{l=1}^{m} k\left(\sum_{i=1}^{s}\sum_{j=1}^{m} p_{(k-1)m^2s-(l-1)ms+(i-1)m+j}\right)+$$

$$+\sum_{k=1}^{nm}\sum_{i=1}^{s}\sum_{j=1}^{m} i p_{(k-1)ms+(i-1)m+j} \qquad (3)$$

Нами знайдений явний вираз метематичного сподівання кількості паралельних порівнянь у випадку різних законів розподілу ймовірностей звертання до записів. Та обчислені оптимальні розміри блоків за яких математичне сподівання досягає мінімуму.

II. Оптимальні стратегії

Використовуючи метод *m*-паралельного послідовного перегляду нами побудовано оптимальні стратегії паралельного пошуку інформації у послідовних упорядкованих файлах баз даних, що зберігається у зовнішній пам'яті ЕОМ, до складу якого входять *m* процесорів, які працюють паралельно і мають спільне поле пам'яті.

Припустимо, що файл, який містить *N* записів, поділений на *n* блоків, у кожному з яких є *ml* записів. Нехай $a_0 = b_0 + d_0 ml$ – час читання блоку записів в основну пам'ять, де b_0, d_0 – деякі сталі; t_0 – час виконання операції *m*-паралельного послідовного перегляду записів в основній пам'яті; p_i – ймовірність звертання до *i*-го запису файла, E_t – математичне сподівання загального часу, необхідного для пошуку запису у файлі. Приймаємо, що для пошуку запису спочатку відбувається послідовне читання блоків записів в основну пам'ять і їх *m*-паралельний послідовний перегляд. Тоді

$$E_t = \sum_{k=1}^{n}\sum_{i=1}^{l}\sum_{j=1}^{m}\{ka_0 + [(k-1)l + i]t_0\}\times$$

$$\times p_{(k-1)ml+(i-1)m+j}. \qquad (4)$$

Нами знайдений явний вираз метематичного сподівання загального часу, необхідного для пошуку записів у файлі, у випадку різних законів розподілу ймовірностей звертання до записів. Та знайдені оптимальні значення параметрів *n* та *l* за яких математичне сподівання досягає мінімуму.

Висновок

Розглянуто метод *m*-паралельного послідовного перегляду та два варіанти методу *m*-паралельного блочного пошуку. Досліджено ефективність цих методів для таких законів розподілу ймовірностей звертання до записів як: рівномірний, «бінарний», Зіпфа й узагальнений, частковим випадком якого є розподіл, що наближено задовольняє правило «80-20». За критерій ефективності взято математичне сподівання кількості паралельних порівнянь, необхідних для пошуку запису у файлі. Проведено порівняльний аналіз ефективності методів і для кожного розглянутого закону розподілу ймовірностей звертання до записів визначено свій найкращий метод.

Побудовано оптимальні стратегії пошуку записів в послідовних файлах, які зберігаються у зовнішній пам'яті багатопроцесорної ЕОМ, для різних законів розподілу ймовірностей звертання до записів. За критерій ефективності взято математичне сподівання загального часу, необхідного для пошуку запису у файлі. Визначені значення параметрів, за яких математичне сподівання досягає мінімуму.

[1] Лісовець В. Я., Цегелик Г. Г. Метод *m*-паралельного послідовного перегляду записів та його використання для пошуку інформації у послідовних файлах баз даних // *Фізико-математичне моделювання та інформаційні технології*. – 2007. – Вип. 5. — С. 109-119.

[2] Лісовець В., Цегелик Г. Метод *m*-паралельного послідовного пошуку записів у файлах баз даних і його ефективність // *Вісн. Львів. ун-ту. Сер. прикл. матем. та інформ.* –2006. – Вип. 13.– С. 177-186.

[3] Лісовець В. Я., Цегелик Г. Г. Метод *m*-паралельного блочного пошуку записів у файлах баз даних та його ефективність // *Відбір та обробка інформації*. – 2007. – Вип. 27(103). – С. 87-92.

[4] Лісовець В. Я., Цегелик Г. Г. Один з варіантів методу *m*-паралельного блочного пошуку записів і його ефективність // *Фізико-математичне моделювання та інформаційні технології*. – 2008. – Вип. 7. – С. 103-111.

[5] Кнут Д. *Искусство программирования для ЭВМ. Т. 3: Сортировка и поиск*. – М.: Изд. дом „Вильямс", 2000. – 832 с.

[6] Мартин Дж. *Организация баз данных в вычислительных системах*. – М: Мир, 1980. – 644 с.

[7] Цегелик Г. Г. *Организация и поиск информации в базах данных*. – Львов: Вища шк., 1987. – 176 с.

[8] Цегелик Г.Г. *Системы распределенных баз данных*. – Львов: Світ, 1990, – 168с.