УДК 519.68

А. В. Мельничин, асп.;
М. І. Філяк,
Г. Г. Цегелик , д. ф.-м. н., проф.

ЕФЕКТИВНІСТЬ МЕТОДУ ПОШУКУ ІНФОРМАЦІЇ У ФАЙЛАХ БАЗ ДАНИХ, ЯКИЙ ВРАХОВУЄ РОЗПОДІЛ ІМОВІРНОСТЕЙ ЗВЕРТАННЯ ДО ЗАПИСІВ

Запропоновано метод пошуку інформації у файлах баз даних, який враховує розподіл імовірностей звертання до записів. Досліджено ефективність цього методу в порівнянні з методами послідовного перегляду, двійкового та блочного пошуку для різних законів розподілу ймовірностей звертання до записів.

Вступ

Найвживанішими методами пошуку інформації у файлах баз даних на сучасному етапі розвитку інформаційних систем є метод послідовного перегляду, однорівневий та багаторівневий блочний і двійковий пошук. У [1—11] досліджено ефективність цих методів для різних законів розподілу ймовірностей звертання до записів (рівномірного і «бінарного», закону Зіпфа, узагальненого закону, окремим випадком якого є розподіл, що наближено задовольняє правило «80-20»). За критерій ефективності прийнято математичне сподівання кількості порівнянь, необхідних для пошуку запису у файлі. Для кожного закону розподілу ймовірностей проведено порівняльний аналіз ефективності методів і визначено свій найкращий метод пошуку. Однак алгоритми методів послідовного перегляду і двійкового пошуку ніяк не враховують розподіл імовірностей звертання до записів, а алгоритми методів однорівневого та багаторівневого блочного пошуку розподіл імовірностей звертання до записів враховують лише при розбитті записів файла на блоки та підблоки однакової довжини у випадку побудови оптимальних схем блочного пошуку. Тому постає задача побудови такого методу пошуку, який би суттєво враховував розподіл імовірностей звертання до записів. Якраз такий метод пошуку і пропонується в даній роботі. Крім того, ефективність цього методу досліджується в порівнянні з методами послідовного перегляду, блочного з оптимальним розміром блоків та двійкового пошуку для згаданих законів розподілу ймовірностей звертання до записів.

Постановка задачі

Розглянемо послідовний упорядкований файл, записи якого характеризуються значеннями деякого ключа. Нехай N — кількість записів файла, p_i — ймовірність звертання до i-го запису файла, K_i — значення ключа, яким характеризується i-й запис файла. Треба побудувати такий метод пошуку записів у файлі, алгоритм якого суттєво враховував би розподіл імовірностей звертання до записів, та дослідити ефективність цього методу в порівнянні з ефективністю методів послідовного перегляду, двійкового та блочного з оптимальним розміром блоків пошуку для різних відомих законів розподілу ймовірностей звертання до записів.

Алгоритм методу

Для опису алгоритму методу введемо поняття умовно середнього запису серед записів файла. Вважатимемо, що умовно середнім серед записів з порядковими номерами від m до n включно, де $1 \le m < n \le N$, є запис з номером r, якщо

$$\min_{m \le k \le n} \left| \sum_{i=m}^{k-1} p_i - \sum_{i=k+1}^{n} p_i \right|$$

досягається для $k = r$. Якщо мінімум досягається для двох індексів k, то за r приймаємо менший із

них. Крім того, якщо $k = m$ і $k = n$ суми $\sum_{i=m}^{k-1} p_i$ і $\sum_{i=k+1}^{n} p_i$ є невизначеними. Тому вважатимемо, що $\sum_{i=m}^{k-1} p_i = 0$, якщо $k = m$; $\sum_{i=k+1}^{n} p_i = 0$, якщо $k = n$.

Припустимо, що у файлі потрібно знайти запис із значенням ключа K. Алгоритм методу складається із низки кроків. На першому кроці K порівнюється зі значенням ключа запису, який є умовно середнім у файлі. Якщо порівняння успішне (два значення, що порівнюються, збігаються), то на цьому робота алгоритму закінчується. Якщо два значення, що порівнюються, не збігаються, то з порівняння видно, в якій частині файла треба продовжувати пошук. Тоді на другому кроці K порівнюється зі значенням ключа запису, який є умовно середнім у вибраній частині файла. При успішному порівнянні робота алгоритму закінчується, при неуспішному — пошук продовжується у ще меншій частині файла, і т. д. Через скінченну кількість кроків шуканий запис буде знайдений, якщо він міститься у файлі.

Формули для знаходження умовно середнього запису

1. Якщо розподіл ймовірностей звертання до записів є рівномірним, то метод збігається з методом двійкового пошуку. Тоді середнім серед записів з номерами від m до n включно буде запис з номером r, де $r = \lceil (m + n)/2 \rceil$.

2. Нехай ймовірності звертання до записів розподілені за «бінарним» законом, тобто

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

Тоді умовно середнім серед записів з номерами $2k-1, 2k, 2k+1, \ldots, N$ ($k = 1, 2, \ldots, [N/2]$) буде запис з номером $r = 2k$.

3. Припустимо, що ймовірності звертання до записів розподілені за законом Зіпфа, тобто

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

де $H_N = \sum_{k=1}^{N} \frac{1}{k}$ — частинна сума гармонічного ряду. Оскільки

$$\sum_{i=m}^{k-1} p_i - \sum_{i=k+1}^{n} p_i = \frac{1}{H_N}\left(\sum_{i=m}^{k-1} \frac{1}{i} - \sum_{i=k+1}^{n} \frac{1}{i} \right),$$

то умовно середнім серед записів з номерами від m до n включно при $n > m+1$ буде запис з номером $r = k$, де k — індекс, для якого досягається

$$\min_{m \le k \le n} \left| \sum_{i=m}^{k-1} \frac{1}{i} - \sum_{i=k+1}^{n} \frac{1}{i} \right|.$$

При $n = m+1$ умовно середнім буде запис з номером m.

Одержану формулу для знаходження індекса k можна замінити простішою. Справді,

$$\sum_{i=m}^{k-1} \frac{1}{i} = \int_{m}^{k-1} \frac{dx}{x} + \varepsilon_1(k) = \ln(k-1) - \ln m + \varepsilon_1(k);$$

$$\sum_{i=k+1}^{n} \frac{1}{i} = \int_{k+1}^{n} \frac{dx}{x} + \varepsilon_2(k) = \ln n - \ln(k+1) + \varepsilon_2(k),$$

де $\varepsilon_1(k)$ та $\varepsilon_2(k)$ — похибки апроксимації відповідних сум. Тоді

$$\left| \sum_{i=m}^{k-1} \frac{1}{i} - \sum_{i=k+1}^{n} \frac{1}{i} \right| = \left| \ln\left(k^2 - 1\right) - \ln nm + \varepsilon(k) \right|,$$

де $\varepsilon(k) = \varepsilon_1(k) - \varepsilon_2(k)$. Тому для відшукання k можна використати наближену формулу $k \approx \sqrt{nm+1}$. Оскільки k повинно бути цілим числом, то можемо прийняти $k = \lceil \sqrt{nm+1} \rceil$.

4. Якщо ймовірності звертання до записів задовольняють узагальнений закон розподілу, тобто

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

де $0 < c < 1$, $H_N^{(c)} = \sum_{k=1}^{N} \frac{1}{k^c}$ — частинна сума узагальненого гармонічного ряду, то умовно середнім серед записів з номерами від m до n включно при $n > m+1$ буде запис з номером $r = k$, де k — індекс, для якого досягається

$$\min_{m \le k \le n} \left| \sum_{i=m}^{k-1} \frac{1}{i^c} - \sum_{i=k+1}^{n} \frac{1}{i^c} \right|.$$

При $n = m+1$ умовно середнім буде запис з номером m. Оскільки

$$\sum_{i=m}^{k-1} \frac{1}{i^c} = \int_{m}^{k-1} x^{-c} dx + \varepsilon_1^{(c)}(k) = \frac{1}{1-c}\left((k-1)^{1-c} - m^{1-c}\right) + \varepsilon_1^{(c)}(k);$$

$$\sum_{i=k+1}^{n} \frac{1}{i^c} = \int_{k+1}^{n} x^{-c} dx + \varepsilon_2^{(c)}(k) = \frac{1}{1-c}\left(n^{1-c} - (k+1)^{1-c}\right) + \varepsilon_2^{(c)}(k),$$

де $\varepsilon_1^{(c)}(k)$ і $\varepsilon_2^{(c)}(k)$ — похибки апроксимації відповідних сум, то

$$\left| \sum_{i=m}^{k-1} \frac{1}{i^c} - \sum_{i=k+1}^{n} \frac{1}{i^c} \right| = \left| \frac{1}{1-c}\left((k-1)^{1-c} + (k+1)^{1-c} - m^{1-c} - n^{1-c}\right) + \varepsilon^{(c)}(k) \right|,$$

де $\varepsilon^{(c)}(k) = \varepsilon_1^{(c)}(k) - \varepsilon_2^{(c)}(k)$.

Із умови

$$(k-1)^{1-c} + (k+1)^{1-c} = m^{1-c} + n^{1-c}$$

або

$$k^{1-c}\left(1 - \frac{1}{k}\right)^{1-c} + k^{1-c}\left(1 + \frac{1}{k}\right)^{1-c} = m^{1-c} + n^{1-c}$$

отримуємо таку наближену формулу для знаходження k

$$k \approx \left(\frac{1}{2}\left(m^{1-c} + n^{1-c}\right)\right)^{\frac{1}{1-c}}.$$

Оскільки k повинно бути цілим числом, то можемо прийняти

$$k = \left[\left(\frac{1}{2}\left(m^{1-c} + n^{1-c}\right)\right)^{\frac{1}{1-c}}\right].$$

Зокрема, при $c = 0$ (тобто у випадку рівномірного розподілу ймовірностей) із отриманої формули дістаємо $k = \left[(m+n)/2\right]$.

Математичне сподівання кількості порівнянь, необхідних для пошуку запису

У випадку рівномірного розподілу ймовірностей звертання до записів середня кількість порівнянь, необхідних для пошуку запису у файлі, виражається формулою [12]

$$E = l - \frac{2^l - l - 1}{N},$$

де $l = 1 + \left[\log_2 N\right]$.

Якщо ймовірності звертання до записів задовольняють «бінарний» закон розподілу, то для математичного сподівання кількості порівнянь, необхідних для пошуку запису у файлі, отримуємо вираз

$$E = \frac{1}{2^2} + \sum_{i=2}^{k} i\left(\frac{1}{2^{2i-3}} + \frac{1}{2^{2i}}\right) + \frac{3k+2}{2^{2k}}$$

при $N = 2k$ і

$$E = \frac{1}{2^2} + \sum_{i=2}^{k} i\left(\frac{1}{2^{2i-3}} + \frac{1}{2^{2i}}\right) + \frac{3k+3}{2^{2k}}$$

при $N = 2k+1$.

Для знаходження математичного сподівання кількості порівнянь, необхідних для пошуку запису у файлі, у разі закону Зіпфа та узагальненого розподілу будемо користуватись алгоритмом методу.

Порівняльна ефективність методу

У таблиці показаний розрахунок математичного сподівання кількості порівнянь, необхідних для пошуку запису у файлі, для різних законів розподілу ймовірностей і деяких N у випадку методу, який враховує розподіл імовірностей звертання до записів, а також у разі методів послідовного перегляду, двійкового та блочного з оптимальним розміром блоків пошуку.

Математичне сподівання кількості порівнянь, необхідних для пошуку запису у файлі

N	Методи пошуку	Закон розподілу						
		$c = 0$	$c = 0{,}2$	$c = 0{,}4$	$c = 0{,}6$	$c = 0{,}8$	Зіпфа ($c = 1$)	«Бінарний»
127	1	6,06	6,08	6,00	5,78	5,43	4,90	2,00
	2	64,00	57,36	49,65	41,02	31,99	23,43	2,00
	3	6,05	43,14	43,45	44,03	44,97	46,33	52,24
	4	12,27	11,53	10,58	9,47	8,21	6,89	2,67
2047	1	10,01	10,02	9,88	9,47	8,67	7,41	2,00
	2	1024,00	911,20	773,24	607,47	422,93	249,60	2,00
	3	10,01	683,39	684,77	689,00	699,76	721,15	835,76
	4	46,24	43,43	39,65	34,47	27,68	20,00	2,67
8191	1	12,00	12,02	11,86	11,39	10,36	8,59	2,00
	2	2048,00	1821,60	1542,96	1203,90	820,55	460,40	2,00
	3	11,01	1366,15	1368,07	1374,79	1394,16	1436,28	1671,53
	4	91,50	86,05	78,61	68,00	53,22	35,77	2,67
16383	1	13,00	13,01	12,85	12,35	11,21	9,17	2,00
	2	8192,00	7283,37	6156,22	4757,61	3129,37	1593,52	2,00
	3	13,00	5462,36	5466,01	5482,71	5544,56	5706,81	6686,11
	4	129,00	121,37	110,94	95,86	74,27	48,22	2,67

Примітки. 1 – метод, який враховує розподіл імовірностей звертання до записів; 2 – метод послідовного перегляду; 3 – метод двійкового пошуку; 4 – метод блочного пошуку з оптимальним розміром блоків.

Зауважимо, що при обчисленні математичного сподівання кількості порівнянь, необхідних для пошуку запису у файлі, для різних законів розподілу ймовірностей звертання до записів у випадку методу послідовного перегляду нами використані формули для математичного сподівання, наведені в [7]; для обчислення математичного сподівання у випадку методу двійкового пошуку ми скористались формулою [4]

$$E = \sum_{i=1}^{l} \sum_{k=1}^{2^{i-1}} i p_{(2k-1)n_i},$$

яка справджується при $N = 2^l - 1$, де l – будь-яке натуральне число $(l \geq 2)$, $n_i = m/2^{i-1}$, $m = [N/2] + 1$ (в загальному випадку можна скористатись алгоритмом методу), а для обчислення математичного сподівання у разі методу блочного пошуку з оптимальним розміром блоків використані формули [3].

Висновки

Побудовано метод пошуку записів у файлах баз даних, який суттєво враховує розподіл імовірностей звертання до записів. Для порівняння ефективності побудованого методу і методів послідовного перегляду, блочного з оптимальним розміром блоків та двійкового пошуку для всіх розглянутих законів розподілу ймовірностей звертання до записів проведено розрахунок математичного сподівання кількості порівнянь, необхідних для пошуку запису у файлі, для різної кількості записів *N*. Як видно з таблиці, побудований метод за ефективністю значно переважає згадані методи для всіх розглянутих законів розподілу ймовірностей звертання до записів, крім рівномірного для методу двійкового пошуку і «бінарного» для методу послідовного перегляду та блочного пошуку з оптимальним розміром блоків. Метод, який враховує розподіл імовірностей звертання до записів, у випадку рівномірного розподілу ймовірностей, збігається з методом двійкового пошуку, а у випадку «бінарного» розподілу — з методом послідовного перегляду.

СПИСОК ЛІТЕРАТУРИ

1. Кнут Д. Искусство программирования для ЭВМ. Т. 3: Сортировка и поиск. — М.: Издательский дом «Вильямс», 2000. — 832 с.

2. Мартин Дж. Организация баз даных в вычислительных системах. — М.: Мир, 1980. — 644 с.

3. Мельничин А. В., Цегелик Г. Г. Аналіз методів пошуку інформації в файлах баз даних для різних законів розподілу ймовірностей звертання до записів // Комп'ютерні технології друкарства. — 2006. — № 15. — С. 86—96.

4. Мельничин А. В., Цегелик Г. Г. Ефективність методу двійкового пошуку інформації у файлах баз даних для різних законів розподілу ймовірностей звертання до записів // Вісник Львівського університету. Сер. прикл. математика та інформатика. — 2006. — Вип. 11. — С. 213 — 218.

5. Філяк М. І., Цегелик Г. Г. Ефективність методів послідовного перегляду і блочного пошуку для різних законів розподілу ймовірностей звертання до записів // Вісник Львівського. університету. Сер. мех.-мат. —1998. — Вип. 50. — С. 200—203.

6. Філяк М. І., Цегелик Г. Г. Ефективність методу дворівневого блочного пошуку у впорядкованих файлах для різних законів розподілу ймовірностей звертання до записів // Вісник Львівського університету. Сер. прикл. математика та інформатика. — 1999. — Вип. 1. — С. 227—230.

7. Філяк М. І., Цегелик Г. Г., Дороцька Х. С. Порівняльний аналіз ефективності методу послідовного перегляду для різних законів розподілу ймовірностей звертання до записів // Вісник НУ «Львівська політехніка». Сер. Інформаційні системи та мережі. — 2000. — № 406. — С. 226—231.

8. Філяк М. І., Цегелик Г. Г. Метод r-рівневого блочного пошуку записів у впорядкованих файлах і його ефективність // Вісник Львівського університету. Сер. прикл. математика та інформатика. — 2000. — Вип. 3. — С. 169—173.

9. Цегелик Г. Г. Мельничин А. В. Ефективність методу r-рівневого блочного пошуку для різних законів розподілу ймовірностей звертання до записів // Вісник НУ «Львівська політехніка». Сер. Інформаційні системи та мережі. — 2005. — № 549. — С. 184—192.

10. Цегелик Г. Г., Філяк М. І. Про ефективність методу r-рівневого блочного пошуку записів у впорядкованих файлах // Вісник Львівського університету. Сер. прикл. математик та інформатика. — 2002. — Вип. 5. — С. 174 —177.

11. Цегелик Г. Г., Філяк М. І., Дороцька Х. С. Порівняльний аналіз ефективності методу блочного пошуку для різних законів розподілу ймовірностей звертання до записів // Комп'ютерні технології друкарства. — 2000. — № 5. — С. 320—326.

12. Цегелик Г. Г. Методы автоматической обработки информации. — Львов: Вища школа. 1981. — 132 с.

Мельничин Андрій Володимирович — аспірант, ***Філяк Марія Іванівна*** — науковий співробітник, ***Цегелик Григорій Григорович*** — завідувач кафедри.

Кафедра математичного моделювання соціально-економічних процесів, Львівський національний університет імені Івана Франка