УДК 519.68

Метод пошуку інформації у файлах баз даних, який враховує розподіл імовірностей звертання до записів

Григорій Цегелик[1], Андрій Мельничин[2]

[1] д. ф.-м. н., професор, Львівський національний університет імені Івана Франка, вул. Університетська, 1, Львів, 79000, e-mail: kafmmsep@franko.lviv.ua
[2] Львівський національний університет імені Івана Франка, вул. Університетська, 1, Львів, 79000, e-mail: andrue_m@mail333.com

Пропонується метод пошуку інформації у файлах баз даних, який враховує розподіл імовірностей звертання до записів, в основі якого лежить поняття умовно середнього запису. Виводяться формули для визначення умовно середнього запису у випадку різних законів розподілу ймовірностей. Досліджується ефективність цього методу порівняно з методами послідовного перегляду та двійкового пошуку для таких законів розподілу ймовірностей як рівномірний, «бінарний», Зіпфа, узагальнений, частковим випадком якого є розподіл, що наближено задовольняє правило «80-20». За критерій ефективності прийнято математичне сподівання кількості порівнянь, необхідних для пошуку запису у файлі.

Ключові слова: методи пошуку, файли баз даних, закони розподілу ймовірностей.

Вступ. Найуживанішими методами пошуку інформації у файлах баз даних є метод послідовного перегляду, однорівневий і багаторівневий блочний і двійковий пошук. Дослідженню ефективності цих методів для різних законів розподілу ймовірностей звертання до записів присвячена низка праць. Зокрема, в [1] розглянуто ефективність методу послідовного перегляду для таких законів розподілу ймовірностей як рівномірний, «бінарний», Зіпфа та правило «80-20». У [2] побудовано оптимальну схему блочного пошуку для рівномірного розподілу ймовірностей. Повне та всестороннє дослідження ефективності методів послідовного перегляду, однорівневого та багаторівневого блочного і двійкового пошуку для різних законів розподілу ймовірностей звертання до записів (рівномірного, «бінарного», Зіпфа, узагальненого, частковим випадком якого є розподіл, що наближено задовольняє правило «80-20») проведено в [3-11]. 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 найкращий метод пошуку. Однак алгоритми методів послідовного перегляду та двійкового пошуку 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 файлу на блоки та підблоки однакової довжини у випадку побудови оптимальних схем блочного пошуку. Тому постає задача побудови такого методу

пошуку, який би враховував розподіл імовірностей звертання до записів. Саме такий метод пошуку і пропонується в даній роботі. Досліджується також ефективність цього методу порівняно з методами послідовного перегляду та двійкового пошуку для згаданих законів розподілу ймовірностей звертання до записів.

1. 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-ий запис файла. Треба побудувати такий метод пошуку записів у файлі, алгоритм якого враховував би розподіл імовірностей звертання до записів, та дослідити ефективність цього методу порівняно з ефективністю методів двійкового пошуку та послідовного перегляду для різних відомих законів розподілу ймовірностей звертання до записів.

2. Алгоритм методу

Для опису алгоритму методу введемо поняття умовно середнього запису серед записів файла. Вважатимемо, що умовно середнім серед записів із порядковими номерами від m до n, де $1 \le m < n \le N$, є запис із 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 значень індексу k, то за r приймаємо менший із них. Зазначимо також, що якщо $k = m$ або $k = n$, то суми $\sum_{i=m}^{k-1} p_i$ або ж $\sum_{i=k+1}^{n} p_i$ є невизначеними. Тому приймаємо, що $\sum_{i=m}^{k-1} p_i = 0$, якщо $k = m$; $\sum_{i=k+1}^{n} p_i = 0$, якщо $k = n$.

Нехай у файлі потрібно знайти запис із значенням ключа K. Алгоритм методу складається з низки кроків. На першому кроці K порівнюється зі значенням ключа запису, який є умовно середнім у файлі. Якщо порівняння успішне (два значення, що порівнюються, співпадають), то на цьому робота алгоритму завершується. Якщо значення, що порівнюються, не співпадають, то з їх порівняння бачимо, в якій частині файла треба продовжувати пошук. Тоді, на наступному кроці K порівнюється зі значенням ключа запису, який є умовно середнім у вибраній частині файла. У разі успішного порівняння робота алгоритму завершується. При неуспішному — пошук продовжується вже у меншій частині файла. І т. д. Через скінченну кількість кроків шуканий запис буде знайдено, якщо він міститься у файлі.

3. Формули для знаходження умовно середнього запису

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

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

$$p_i = 1/2^i \ \left(i = \overline{1, N-1}\right), \quad p_N = 1/2^{N-1}.$$

Тоді умовно середнім серед записів із номерами $2k-1, 2k, 2k+1, ..., N$ $\left(k = \overline{1, \lfloor N/2 \rfloor}\right)$ буде запис із номером $r = 2k$.

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

$$p_i = \frac{1}{iH_N} \ \left(i = \overline{1, N}\right),$$

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 із номерами від m до n включно ($n > m+1$), буде запис із номером $r = k$, де k — індекс, для якого досягається

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

Для $n = m + 1$ умовно середнім буде запис із номером m.

Одержану формулу для знаходження індексу 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

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

де $\varepsilon(k)=\varepsilon_1(k)-\varepsilon_2(k)$. Тому для відшукання k використаємо наближену формулу $k\approx\sqrt{nm+1}$. Оскільки k повинно бути цілим числом, то приймаємо $k=\left[\sqrt{nm+1}\right]$.

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

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

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

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

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

Із рівності

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

або

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

одержуємо 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 приймаємо

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

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

4. Математичне сподівання кількості порівнянь, необхідних для пошуку запису

Для дослідження ефективності методу за критерій ефективності приймемо математичне сподівання кількості порівнянь, необхідних для пошуку запису у файлі. Запишемо формули математичного сподівання для деяких законів розподілу ймовірностей звертання до записів.

У разі рівномірного розподілу ймовірностей звертання до записів середня кількість порівнянь, необхідних для пошуку запису у файлі, визначається за 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, одержуємо 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 $N = 2k$ та

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

при $N = 2k + 1$.

У разі закону Зіпфа й узагальненого розподілу для знаходження математичного сподівання кількості порівнянь, необхідних для пошуку запису у файлі, будемо застосовувати алгоритм методу, описаний вище.

5. Порівняльна ефективність методу

У табл. 1 із використанням методу, який враховує розподіл імовірностей звертання до записів, приведені результати розрахунку математичного сподівання кількості порівнянь, необхідних для пошуку запису у файлі для різних законів розподілу ймовірностей і деяких *N*. У табл. 2 і табл. 3 для тих же законів розподілу ймовірностей звертання до записів і *N* наведено результати розрахунку математичного сподівання у разі використання методів послідовного перегляду та двійкового пошуку відповідно.

Зауважимо, що у табл. 2 при обчисленні математичного сподівання кількості порівнянь використані формули для математичного сподівання, записані в [7]. Для обчислення математичного сподівання у випадку методу двійкового пошуку (див. табл. 3) використано формулу [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 $(l \geq 2)$, $n_i = m / 2^{i-1}$, $m = [N/2] + 1$ (у загальному випадку можна скористатися алгоритмом методу).

Таблиця 1

Математичне сподівання кількості порівнянь, необхідних для пошуку запису у файлі. Метод із врахуванням розподілу ймовірностей звертання до записів

N	**Закон розподілу**						
	c=0	*c*=0,2	*c*=0,4	*c*=0,6	*c*=0,8	Зіпфа (*c*=1)	«Бінарний»
3	1,667	1,674	1,685	1,697	1,711	1,727	1,750
7	2,429	2,442	2,471	2,423	2,381	2,264	1,984
15	3,267	3,306	3,313	3,219	3,123	2,926	2,000
31	4,161	4,199	4,159	4,055	3,847	3,655	2,000
63	5,095	5,131	5,068	4,900	4,630	4,265	2,000
127	6,055	6,082	6,004	5,779	5,431	4,901	2,000
255	7,031	7,055	6,951	6,678	6,210	5,554	2,000
511	8,018	8,039	7,918	7,598	7,033	6,164	2,000
1023	9,010	9,027	8,894	8,529	7,843	6,814	2,000
2047	10,005	10,021	9,877	9,471	8,673	7,413	2,000
4095	11,003	11,018	10,866	10,426	9,502	7,999	2,000
8191	12,002	12,016	11,858	11,386	10,355	8,589	2,000
16383	13,001	13,014	12,852	12,352	11,205	9,174	2,000
32767	14,000	14,013	13,847	13,326	12,075	9,755	2,000

Таблиця 2

Математичне сподівання кількості порівнянь, необхідних для пошуку запису у файлі. Метод послідовного перегляду

N	**Закон розподілу**						
	c=0	*c*=0,2	*c*=0,4	*c*=0,6	*c*=0,8	Зіпфа (*c*=1)	«Бінарний»
3	2	1,93	1,85	1,78	1,71	1,79	1,750
7	4	3,75	3,49	3,22	2,96	2,77	1,984
15	8	7,36	6,67	5,95	5,22	4,57	2,000
31	16	14,53	12,91	11,17	9,40	7,73	2,000
63	32	28,81	25,23	21,28	17,21	13,35	2,000
127	64	57,36	49,65	41,02	31,99	23,43	2,000
255	128	114,35	98,21	79,77	60,19	41,68	2,000
511	256	288,25	194,96	156,20	114,39	75,00	2,000
1023	512	455,94	387,94	307,43	219,20	136,26	2,000
2047	1024	911,20	773,24	607,47	422,93	249,60	2,000
4095	2048	1821,60	1542,96	1203,90	820,55	460,40	2,000
8191	4096	3642,25	3081,23	2391,24	1599,36	854,32	2,000
16383	8192	7283,37	6156,22	4757,61	3129,37	1593,52	2,000
32767	16384	14565,40	12304,20	9477,81	6142,71	2985,83	2,000

Таблиця 3

Математичне сподівання кількості порівнянь, необхідних для пошуку запису у файлі. Метод двійкового пошуку

N	**Закон розподілу**						
	c=0	*c*=0,2	*c*=0,4	*c*=0,6	*c*=0,8	Зіпфа (*c*=1)	«Бінарний»
3	1,45	1,67	1,68	1,70	1,71	1,73	1,75
7	2,35	3,02	3,05	3,09	3,14	3,19	3,28
15	3,24	5,71	5,77	5,87	5,98	6,10	6,50
31	4,15	11,07	11,19	11,36	11,62	11,93	13,06
63	5,09	21,77	21,97	22,30	22,79	23,45	26,12
127	6,05	43,14	43,45	44,03	44,97	46,33	52,24
255	7,03	85,86	86,32	87,30	89,07	91,80	104,47
511	8,02	171,25	171,93	173,54	176,80	182,26	208,94
1023	9,01	341,98	342,96	345,58	351,53	362,36	417,88
2047	10,01	683,39	684,77	689,54	699,76	721,15	835,76
4095	11,01	1366,15	1368,07	1374,79	1394,16	1436,28	1671,53
8191	12,01	2731,58	2734,24	2744,87	2779,55	2862,29	3343,05
16383	13,00	5462,36	5466,01	5482,71	5544,56	5706,81	6686,11
32767	14,00	10923,8	10928,03	10954,90	11064,90	11362,60	13372,20

Висновки. Запропоновано метод пошуку записів у файлах баз даних, який враховує розподіл імовірностей звертання до записів. Для порівняння ефективності побудованого методу і методів послідовного перегляду та двійкового пошуку для розглянутих законів розподілу ймовірностей звертання до записів проведено розрахунок математичного сподівання кількості порівнянь, необхідних для пошуку запису у файлі, для різної кількості записів N. Побудований метод за ефективністю значно переважає методи двійкового пошуку та послідовного перегляду для всіх розглянутих законів розподілу ймовірностей звертання до записів, окрім рівномірного для методу двійкового пошуку та «бінарного» для методу послідовного перегляду. Метод, який враховує розподіл імовірностей звертання до записів, у випадку рівномірного розподілу ймовірностей співпадає з методом двійкового пошуку, а у випадку «бінарного» розподілу — з методом послідовного перегляду.

Література

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

[2] *Мартин Дж.* Организация баз данных 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. — № 16. — С. 220-236.

[4] *Мельничин А. В., Цегелик Г. Г.* Ефективність методу двійкового пошуку інформації у файлах баз даних для різних законів розподілу ймовірностей звертання до записів // Вісн. Львів. ун-ту. 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 // Вісн. Львів. ун-ту. 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 // Вісн. Львів. ун-ту. 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 // Вісн. 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 // Вісн. Львів. ун-ту. 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 // Вісн. НУ «Львівська політехніка». Сер. інформаційні системи та мережі. — 2006. — № 415. — С. 211-221.

[10] *Цегелик Г. Г., Філяк М. І.* Про ефективність методу r-рівневого блочного пошуку записів у впорядкованих файлах // Вісн. Львів. ун-ту. Сер. прикл. математика 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 шк., 1981. — 132 с.

The Method of the Information Search in Database Files, which Considers the Probability Distribution of Requests to Records

Hryhoriy Tsehelyk, Andriy Melnytchyn

The method of the information search in database files, which considers the probability distribution of request to records, has been constructed. Formulas for identification of record, which is located in the middle of the file, under certain conditions have been proposed for different laws of distribution. The comparative analysis of efficiency of method with the linear search method and binary search method for different laws of probability distribution of requests to records has been investigated. The mathematical expectation of number of comparisons was used as an efficiency criterion.

Метод поиска информации в файлах баз данных, учитывающий распределение вероятностей обращения к записям

Григорий Цегелик, Андрей Мельничин

Предлагается метод поиска информации в файлах баз данных, учитывающий распределение вероятностей обращения к записям, в основании которого лежит понятие условно средней записи. Выведены формулы для определения условно средней записи для разных законов распределения вероятностей. Исследуется эффективность метода по сравнению с методами последовательного пересмотра и двоичного поиска для таких законов распределения вероятностей обращения к записям, как равномерный, «бинарный», Зипфа, обобщенный, частным случаем которого является распределение, приближенно удовлетворяющее правило «80-20». За критерий эффективности принято математическое ожидание количества сравнений, необходимых для поиска записи в файле.

Отримано 22.03.06