УДК 519.68

Використання методу найменших квадратів для побудови наближених методів пошуку інформації у файлах баз даних

Андрій Мельничин[1], Григорій Цегелик[2]

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

Запропоновано новий підхід до побудови наближених методів пошуку інформації у файлах баз даних, який ґрунтується на використанні методу найменших квадратів. Для пошуку записів у файлах будуються апроксимуючі функції, які є лінійними комбінаціями систем функцій Чебишева на відповідних проміжках. Вибираючи різним чином системи функцій Чебишева, отримуємо різні апроксимації. За такого підходу наближені методи враховують тільки розподіл значень ключа та не враховують розподіл ймовірностей звертання до записів. Ефективність даного підходу досліджується на реальних файлах і порівнюється з методами послідовного перегляду, блочного з оптимальним розміром блоків і двійкового пошуків. За критерій ефективності прийнято середню кількість порівнянь, необхідних для пошуку запису у файлі.

Ключові слова: методи пошуку, файли баз даних, метод найменших квадратів, система функцій Чебишева.

Вступ. Основний акцент під час розв'язування різноманітних задач із використанням концепції баз даних переноситься з процедур обробки даних на процедури організації збереження та пошуку інформації у файлах баз даних. Тому продуктивність обчислювальних систем, орієнтованих на роботу з величезними базами даних, значною мірою визначається ефективністю методів пошуку інформації у файлах баз даних.

Для пошуку записів у файлах баз даних, зазвичай, використовують точні методи. Найуживанішим серед них є метод послідовного перегляду, однорівневий і багаторівневий блочний та двійковий пошуки [1-6]. Ефективність цих методів залежить від закону розподілу ймовірностей звертання до записів і не залежить від розподілу значень ключа, яким характеризуються записи файла. Тому виникає потреба мати такий метод пошуку, який би суттєво враховував розподіл значень ключа й ефективність якого не залежала б від розподілу ймовірностей звертання до записів. Саме такий метод і пропонується в даній роботі.

1. Постановка задачі

Розглянемо послідовний упорядкований файл, який містить N записів. Нехай K_i $(i = \overline{1, N})$ — значення ключа, яким характеризується i-й запис файла. Позначимо

$x_i = i, y_i = K_i,$ а через $y = K(x)$ — функцію дискретного аргументу, визначену на множині натуральних чисел від 1 до N: $K(x_i) = y_i$. Для функції $y = K(x)$ на проміжку $[1, N]$ побудуємо апроксимуючу функцію $y = \varphi(x)$ таку, щоб величина $\sum_{i=1}^{N}\left[y_i - \varphi(x_i)\right]^2$ досягала мінімуму. Для забезпечення єдиності розв'язку функцію $\varphi(x)$ означимо як лінійну комбінацію функцій Чебишева $\varphi_0(x), \varphi_1(x), ..., \varphi_m(x)$ на проміжку $[1, N]$, $m < N$ [7]

$$\varphi(x) = \sum_{k=0}^{m} a_k \varphi_k(x),$$

Отже, нам потрібно знайти коефіцієнти $a_0, a_1, ..., a_m$, для яких величина

$$F(a_0, a_1, ..., a_n) = \sum_{i=1}^{N}\left[y_i - \sum_{k=0}^{m} a_k \varphi_k(x_i)\right]^2$$

досягає мінімуму. Для знаходження цих коефіцієнтів одержуємо таку систему лінійних алгебраїчних рівнянь

$$\frac{\partial F}{\partial a_j} = 0, \quad j = \overline{0, m},$$

або

$$\sum_{i=1}^{N}\left[y_i - \sum_{k=0}^{m} a_k \varphi_k(x_i)\right]\varphi_j(x_i) = 0, \quad j = \overline{0, m}.$$

Перепишемо систему в такому вигляді

$$\sum_{k=0}^{m} a_k \sum_{i=1}^{N} \varphi_k(x_i)\varphi_j(x_i) = \sum_{i=1}^{N} y_i \varphi_j(x_i), \quad j = \overline{0, m}.$$

Якщо ввести позначення

$$\alpha_{jk} = \sum_{i=1}^{N} \varphi_k(x_i)\varphi_j(x_i), \quad \beta_j = \sum_{i=1}^{N} y_i \varphi_j(x_i),$$

то система рівнянь для знаходження коефіцієнтів a_k $\left(k = \overline{0, m}\right)$ прийме вигляд

$$\sum_{k=0}^{m} \alpha_{jk} a_k = \beta_j, \quad j = \overline{0, m}.$$

Вибираючи різним чином систему функцій Чебишева $\varphi_0(x), \varphi_1(x), ..., \varphi_m(x)$, одержимо різні апроксимуючі функції.

Якщо за систему функцій Чебишева взяти $\varphi_k(x) = x^k$, $k = \overline{0, m}$, то апроксимуюча функція матиме вигляд $y = \sum_{k=0}^{m} a_k x^k$. Тоді для знаходження значень параметрів a_0, $a_1, ..., a_m$ одержуємо таку систему рівнянь

$$\begin{cases} a_0 N + a_1 \sum_{i=1}^{N} x_i + a_2 \sum_{i=1}^{N} x_i^2 + \ldots + a_m \sum_{i=1}^{N} x_i^m = \sum_{i=1}^{N} y_i; \\ a_0 \sum_{i=1}^{N} x_i + a_1 \sum_{i=1}^{N} x_i^2 + a_2 \sum_{i=1}^{N} x_i^3 + \ldots + a_m \sum_{i=1}^{N} x_i^{m+1} = \sum_{i=1}^{N} y_i x_i; \\ a_0 \sum_{i=1}^{N} x_i^2 + a_1 \sum_{i=1}^{N} x_i^3 + a_2 \sum_{i=1}^{N} x_i^4 + \ldots + a_m \sum_{i=1}^{N} x_i^{m+2} = \sum_{i=1}^{N} y_i x_i^2; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_0 \sum_{i=1}^{N} x_i^m + a_1 \sum_{i=1}^{N} x_i^{m+1} + a_2 \sum_{i=1}^{N} x_i^{m+2} + \ldots + a_m \sum_{i=1}^{N} x_i^{2m} = \sum_{i=1}^{N} y_i x_i^m. \end{cases}$$

Зокрема, при $m = 1$ отримаємо лінійну апроксимацію $y = a_0 + a_1 x$, а при $m = 2$ — квадратичну $y = a_0 + a_1 x + a_2 x^2$.

Якщо за систему функцій Чебишева прийняти $\varphi_k(x) = e^{kx}$, $k = \overline{0, m}$, то апроксимуюча функція матиме вигляд $y = \sum_{k=0}^{m} a_k e^{kx}$. Тоді одержуємо таку систему рівнянь для визначення параметрів a_0, a_1, ..., a_m

$$\begin{cases} a_0 N + a_1 \sum_{i=1}^{N} e^{x_i} + a_2 \sum_{i=1}^{N} e^{2x_i} + \ldots + a_m \sum_{i=1}^{N} e^{mx_i} = \sum_{i=1}^{N} y_i; \\ a_0 \sum_{i=1}^{N} e^{x_i} + a_1 \sum_{i=1}^{N} e^{2x_i} + a_2 \sum_{i=1}^{N} e^{3x_i} + \ldots + a_m \sum_{i=1}^{N} e^{(m+1)x_i} = \sum_{i=1}^{N} y_i e^{x_i}; \\ a_0 \sum_{i=1}^{N} e^{2x_i} + a_1 \sum_{i=1}^{N} e^{3x_i} + a_2 \sum_{i=1}^{N} e^{4x_i} + \ldots + a_m \sum_{i=1}^{N} e^{(m+2)x_i} = \sum_{i=1}^{N} y_i e^{2x_i}; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_0 \sum_{i=1}^{N} e^{mx_i} + a_1 \sum_{i=1}^{N} e^{(m+1)x_i} + a_2 \sum_{i=1}^{N} e^{(m+2)x_i} + \ldots + a_m \sum_{i=1}^{N} e^{2mx_i} = \sum_{i=1}^{N} y_i e^{mx_i}. \end{cases}$$

Зокрема, якщо $m = 1$, то отримаємо експоненційну апроксимацію $y = a_0 + a_1 e^x$.

Побудовані апроксимуючі функції можуть використовуватися для пошуку записів за значеннями ключа.

2. Комп'ютерний експеримент

Нами проведено комп'ютерний експеримент із дослідження ефективності запропонованого підходу на реальних файлах. У таблиці приведені значення ключа файла, на основі якого проводився експеримент. Ці значення були згенеровані випадковим чином.

Нами побудовано лінійну, експоненційну та квадратичну апроксимуючі функції.

Таблиця

Значення ключа файла

X	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*
1	1	21	39	41	77	61	114	81	160
2	3	22	40	42	78	62	115	82	161
3	5	23	44	43	79	63	118	83	165
4	7	24	45	44	80	64	123	84	166
5	10	25	47	45	82	65	124	85	168
6	11	26	48	46	83	66	125	86	169
7	12	27	50	47	84	67	129	87	170
8	18	28	51	48	86	68	130	88	172
9	19	29	53	49	87	69	131	89	173
10	20	30	55	50	90	70	132	90	174
11	21	31	58	51	91	71	135	91	176
12	22	32	60	52	93	72	138	92	177
13	23	33	61	53	94	73	141	93	178
14	25	34	65	54	96	74	144	94	180
15	26	35	67	55	99	75	146	95	182
16	28	36	70	56	100	76	150	96	184
17	29	37	71	57	102	77	151	97	185
18	32	38	72	58	103	78	153	98	186
19	33	39	74	59	106	79	156	99	187
20	35	40	75	60	113	80	157	100	189

$y = 1{,}9483x - 2{,}7691$, $R^2 = 0{,}997$;

$y = 15{,}701e^{0{,}0298x}$, $R^2 = 0{,}7852$;

$y = 0{,}0018x^2 + 1{,}7667x + 0{,}3187$, $R^2 = 0{,}9972$.

Тут R — коефіцієнт кореляції.

Було проведено порівняння ефективності запропонованого підходу з методами послідовного перегляду, блочного з оптимальним розміром блоків і двійкового пошуків. За критерій ефективності приймалася середня кількість порівнянь, необхідних для пошуку запису у файлі. У випадку методу послідовного перегляду, блочного з оптимальним розміром блоків і двійкового пошуків середня кількість порівнянь, необхідних для пошуку запису у файлі, відповідно становить 50,50; 11,00 і 5,78. Якщо пошук здійснювати з використанням лінійної, експоненціальної або квадратичної апроксимацій, тобто відповідно за формулами

$x = \left(y + 2{,}7691\right)/1{,}9483$;

$x = \ln\left[y - \ln(15{,}701)\right]/0{,}0298$;

$x = 23{,}5704\sqrt{433{,}185 + y} - 490{,}754$,

то середня кількість порівнянь, необхідних для пошуку запису у файлі, відповідно становить 1,21; 9,51 та 1,08.

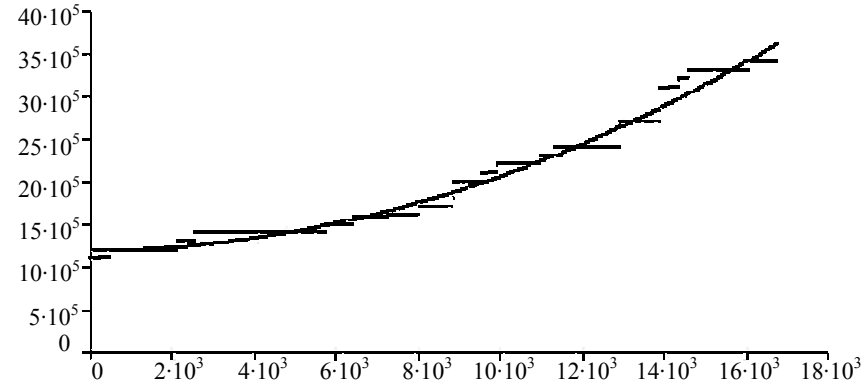

Рис. 1. Розподіл значень ключів та апроксимуюча лінійна функція

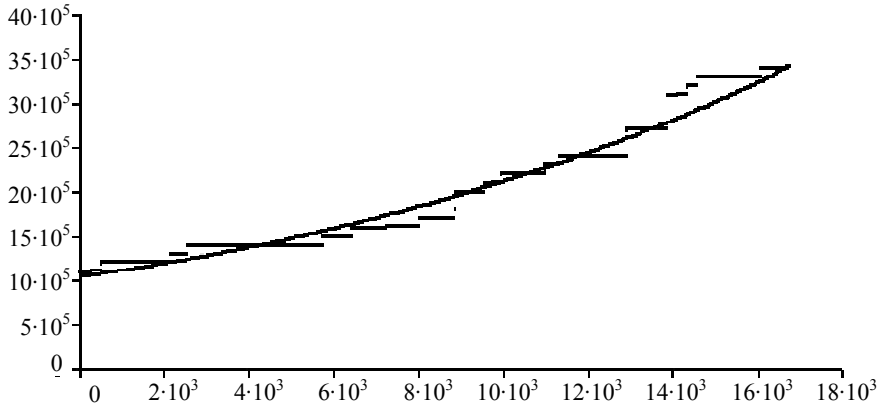

Рис. 2. Розподіл значень ключів й апроксимуюча експоненціальна функція

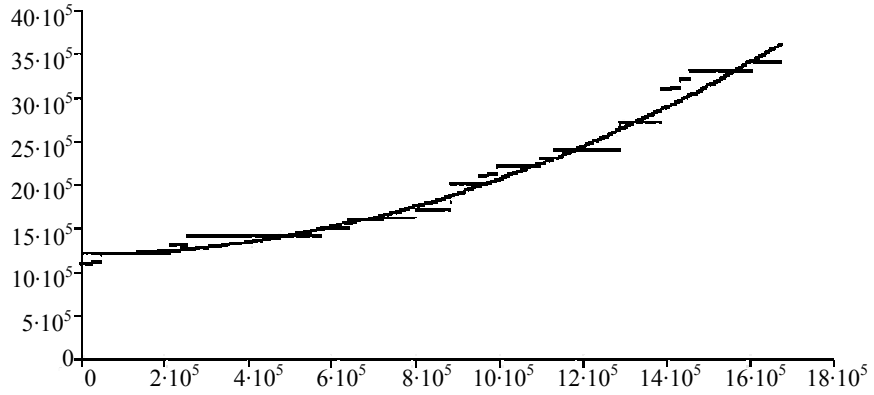

Рис. 3. Розподіл значень ключів та апроксимуюча квадратична функція

Нами також проведено обчислювальний експеримент із файлом реальної бази даних студентів, які навчаються стаціонарно у Львівському національному університеті імені Івана Франка. Цей файл містить 16768 записів. На основі файла значень ключа (номерів залікових книжок) побудовано три апроксимуючі функції: лінійну, квадратичну й експоненційну, а також проведено порівняння ефективності пошуку з використанням цих функцій, з ефективністю методів послідовного перегляду, блочного з оптимальним розміром блоків і двійкового пошуків. За критерій ефективності прийнято середню кількість порівнянь, необхідних для пошуку запису у файлі.

Лінійна апроксимуюча функція (див. рис. 1) є такою $y = 803196 + 144{,}313x$ ($R^2 = 0{,}922$). Для пошуку записів маємо формулу $x = \left(y - 803196\right)/114{,}313$.

Експоненційна апроксимуюча функція (див. рис. 2) є такою $y = \exp(13{,}8572 + 7{,}1889 \cdot 10^{-5}x)$ ($R^2 = 0{,}969$). Пошук записів здійснюється з використанням формули $x = (\ln y - 13{,}8572) / 7{,}1889 \times 10^{-5}$.

Квадратична апроксимуюча функція (див. рис. 3) є такою $y = 0{,}0085x^2 + 1{,}1578x + 1{,}2 \cdot 10^6$ ($R^2 = 0{,}982$), а для пошуку маємо $x = 10{,}86213\sqrt{y - 1099974} - 758{,}94$.

3. Порівняльна ефективність методу

Проведено порівняння ефективності запропонованого підходу з методами послідовного перегляду, блочного з оптимальним розміром блоків і двійкового пошуків. За критерій ефективності прийнято середню кількість порівнянь, необхідних для пошуку запису у файлі. У випадку методу послідовного перегляду, блочного та двійкового пошуків середня кількість порівнянь, необхідних для пошуку запису у файлі, відповідно становить 8384,50; 130,49 і 13,05. Якщо пошук здійснювати з використанням лінійної, експоненціальної та квадратичної апроксимацій, то середня кількість порівнянь, необхідних для пошуку запису у файлі, відповідно є 1251,28; 702,82 та 701,09.

Висновки. Запропоновано метод пошуку у файлах баз даних, що враховує розподіл значень ключа, якими характеризуються записи файла. Проведено порівняння ефективності запропонованого підходу з методами послідовного перегляду, блочного з оптимальним розміром блоків і двійкового пошуків для реальних файлів. Результати порівнянь показали, що запропонований підхід є ефективніший, ніж метод послідовного перегляду, але поступається блочному пошуку з оптимальним розміром блоків і двійковому пошуку.

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., Цегелик Г.* Ефективність методу двійкового пошуку інформації у файлах баз даних для різних законів розподілу ймовірностей звертання до записів // Вісн. Львів. ун-ту. Сер. прикл. матем. та інформ. — 2006. — Вип. 11. — С. 225-229.

[4] *Мельничин А. В., Цегелик Г. Г.* Аналіз методів пошуку інформації в файлах баз даних для різних законів розподілу ймовірностей звертання до записів // Зб. наук. праць Української академії друкарства «Комп'ютерні технології друкарства». — 2006. — № 15. — С. 95-112.

[5] *Цегелик Г. Г., Мельничин А. В.* Порівняльний аналіз ефективності методів пошуку інформації у файлах баз даних // Відбір і обробка інформації. — 2005. — № 23(99). — С. 135-142.

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

[7] *Цегелик Г. Г.* Чисельні методи: підручник. — Львів: Вид. центр Львівського нац. ун-ту ім. І. Франка. — 2004. — 408 с.

Использование метода наименьших квадратов для построения приближенных методов поиска информации в файлах баз данных

Андрей Мельничин, Григорий Цегелик

С использованием метода наименьших квадратов предложен новый подход к построению приближенных методов поиска информации в файлах баз данных. Для поиска записей в файлах строятся аппроксимирующие функции, которые являются линейной комбинацией систем функций Чебышева на соответствующих промежутках. Выбором разных систем функций Чебышева получаем разные аппроксимации. При таком подходе приближенные методы учитывают только распределение значений ключа и не учитывают распределение вероятностей обращения к записям. Эффективность данного подхода исследуется на реальных файлах и сравнивается с методами последовательного пересмотра, блочного с оптимальным размером блоков и двоичного поисков. Критерием эффективности является среднее количество сравнений, необходимое для поиска записи в файле.

Using of least-squares method for creation of approximation methods to information search in database files

Andriy Melnytchyn, Hryhoriy Tsehelyk

A new approximated method of information search in database files is proposed. It is based on the use of the least-squares method. Approximation functions for records search are built. These functions are linear combinations of Chebyshew systems functions on proper intervals. By choosing different systems of Chebyshew functions, we obtain the different approximations. For such approach the approximated methods counts for distribution value of the key only and does not consider probability distribution of requests to the records. The efficiency of proposed approach is investigated by comparison with the linear search method, block search method with the optimum size of blocks and binary search method. The mathematical expectation of number of comparisons was used as an efficiency criterion.

Отримано 01.06.07