УДК 519.95
DOI https://doi.org/10.24144/2616-7700.2019.1(34).102-107

Л. І. Фундак[1], Г. Г. Цегелик[2], М. І. Глебена[3]

[1] Львівський національний університет ім. І. Франка, Львів,
асистент кафедри математичного моделювання соціально-економічних процесів
lfundak@gmail.com
ORCID: https://orcid.org/0000-0001-5091-6971

[2] Львівський національний університет ім. І. Франка, Львів,
професор кафедри математичного моделювання соціально-економічних процесів,
доктор фізико-математичних наук
Hryhoriy.Tsehelyk@gmail.com
ORCID: https://orcid.org/0000-0002-5826-0628

[3] ДВНЗ «Ужгородський національний університет», Ужгород,
доцент кафедри системного аналізу і теорії оптимізації,
кандидат фізико-математичних наук
myroslava.hlebena@uzhnu.edu.ua
ORCID: https://orcid.org/0000-0003-1100-515X

ЕФЕКТИВНІСТЬ МЕТОДУ ДВІЙКОВОГО ПОШУКУ ЗАПИСІВ У ФАЙЛАХ БАЗ ДАНИХ У ВИПАДКУ РОЗПОДІЛУ ЙМОВІРНОСТЕЙ ЗВЕРТАННЯ ДО ЗАПИСІВ ЗА ЗАКОНОМ ЗІПФА

Основний акцент під час розв'язування різноманітних задач з використанням концепції баз даних переноситься з процедур опрацювання інформації на процедури організації збереження та пошуку інформації в базах даних. Тому продуктивність обчислювальних систем, орієнтованих на опрацювання інформації у великих базах даних, головним чином визначається ефективністю методів пошуку інформації у файлах баз даних.

Оскільки в більшості систем опрацювання інформації типовими є випадки нерівномірного розподілу ймовірностей звертання до записів файлів, то дослідження ефективності методів пошуку проводиться для таких типових законів нерівномірного розподілу ймовірностей як бінарний, закон Зіпфа, узагальнений закон.

За критерій ефективності методів приймається математичне сподівання кількості порівнянь, необхідних для пошуку запису у файлі. Деякі часткові результати дослідження ефективності методів пошуку одержані зарубіжними авторами, зокрема вони відображені у монографіях Д. Кнута і Дж. Мартіна. Більш повні дослідження проведені в працях Цегелика Г.Г.

Для пошуку запису у файлі можна використати різні методи: послідовний перегляд; однорівневий чи багаторівневий блоковий пошук; двійковий пошук; метод пошуку, що враховує розподіл ймовірностей звертання до записів; методи пошуку, що використовують індекси тощо. Ефективність цих методів для різних законів розподілу ймовірностей звертання до записів є різною.

Вважатимемо, що файл бази даних упорядкований за зростанням значень ключа. У статті виведено формулу для обчислення математичного сподівання кількості порівнянь, необхідних для пошуку запису у файлі, у випадку розподілу ймовірностей звертання до записів за законом Зіпфа. Зроблено порівняння ефективності методу послідовного перегляду та методу двійкового пошуку у цьому випадку, а також порівняння ефективності двійкового пошуку у випадку рівномірного розподілу ймовірностей і розподілу за законом Зіпфа. На графіках показана залежність математичного сподівання кількості порівнянь від кількості записів у файлі у випадку розподілу ймовірностей звертання до записів за законом Зіпфа.

Ключові слова: розподіл ймовірностей звертання до записів за законом Зіпфа, метод двійкового пошуку, математичне сподівання.

1. Вступ. Основний акцент під час розв'язування різноманітних задач з використанням концепції баз даних (БД) переноситься з процедур опрацювання інформації на процедури організації збереження та пошуку інформації в БД. Тому продуктивність обчислювальних систем, орієнтованих на опрацювання інформації у великих БД, головним чином визначається ефективністю методів пошуку інформації у файлах БД.

Оскільки в більшості систем опрацювання інформації типовими є випадки нерівномірного розподілу ймовірностей звертання до записів файлів [1, 2], то дослідження ефективності методів пошуку проводиться для таких типових законів нерівномірного розподілу ймовірностей як:

– "бінарний"

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

де p_i, $i = 1, 2, \ldots, N$, – ймовірність звертання до i-го запису файлу, N – кількість записів у файлі;

– закон Зіпфа

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

де $H_N = 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{N}$ – частинна сума гармонічного ряду;

– узагальненого закону

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

де $H_N^{(c)} = 1 + \frac{1}{2^c} + \frac{1}{3^c} + \ldots + \frac{1}{N^c}$ – частинна сума узагальненого гармонічного ряду, $0 < c < 1$. При $c = 0,8614$ одержуємо розподіл, який наближено задовольняє правило "80-20".

За критерій ефективності методів приймається математичне сподівання кількості порівнянь, необхідних для пошуку запису у файлі. Деякі часткові результати дослідження ефективності методів пошуку одержані зарубіжними авторами, зокрема вони відображені у монографіях Д. Кнута [1] і Дж. Мартіна [2]. Більш повні дослідження проведені в [3].

2. Основний результат. Для пошуку запису у файлі можна використати різні методи: послідовний перегляд; однорівневий [5] чи багаторівневий блоковий пошук [6]; двійковий пошук; метод пошуку, що враховує розподіл ймовірностей звертання до записів; методи пошуку, що використовують індекси тощо. Ефективність цих методів для різних законів розподілу ймовірностей звертання до записів є різною. Так, у випадку рівномірного розподілу ймовірностей, де $p_i = \frac{1}{N}$, $i = 1, 2, \ldots, N$ найефективнішим методом є метод двійкового пошуку [4]. Максимальна кількість порівнянь для пошуку запису при використанні цього методу рівна

$$k = 1 + [\log_2 N],$$

а середня

$$E = k - \frac{2^k - k - 1}{N}. \tag{1}$$

У випадку нерівномірних законів розподілу ймовірностей звертання до записів формулу для математичного сподівання кількості порівнянь, необхідних для пошуку запису у файлі, можна записати лише у випадку, коли $N = 2^l - 1$, де l – ціле число ($l \geq 2$). Ця формула має вигляд [3]

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

де $n_i = \frac{m}{2^{i-1}}$, $m = \left[\frac{N}{2}\right] + 1$.

Розглянемо файл, який містить N записів. Нехай k_i, $i = 1, 2, \ldots, N$, – значення ключа, яким характеризується i-й запис файлу; p_i, $i = 1, 2, \ldots, N$, – ймовірність звертання до i-го запису файлу. Вважатимемо, що файл упорядкований за зростанням значень ключа. Знайдемо математичне сподівання кількості порівнянь у випадку розподілу ймовірностей звертання до записів за законом Зіпфа. Зробимо порівняння ефективності методу послідовного перегляду та методу двійкового пошуку у випадку розподілу ймовірностей за законом Зіпфа, а також порівняння ефективності двійкового пошуку у випадку рівномірного розподілу ймовірностей і розподілу за законом Зіпфа.

Якщо для пошуку запису у файлі використовувати метод послідовного перегляду, то математичне сподівання кількості порівнянь [3]

$$E_1 = \frac{N}{\ln N + C}. \tag{2}$$

де $C = 0,577\ldots$ – ейлерова стала. Якщо ж для пошуку запису у файлі використовувати метод двійкового пошуку, то математичне сподівання кількості порівнянь обчислюватимемо за формулою [3]

$$E_2 = \sum_{i=1}^{l} i \sum_{k=1}^{2^{i-1}} \frac{1}{H_N (2k-1) n_i} = \frac{1}{H_N} \sum_{i=1}^{l} \frac{i}{n_i} \sum_{k=1}^{2^{i-1}} \frac{1}{2k-1}.$$

Оскільки

$$\sum_{k=1}^{2^{i-1}} \frac{1}{2k-1} = 1 + \frac{1}{3} + \frac{1}{5} + \ldots + \frac{1}{2^i - 1} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots + \frac{1}{2^i - 1} + \frac{1}{2^i} -$$

$$- \left(\frac{1}{2} + \frac{1}{4} + \frac{1}{6} + \ldots + \frac{1}{2^i}\right) = H_{2^i} - \frac{1}{2}\left(1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{2^{i-1}}\right) = H_{2^i} - \frac{1}{2} H_{2^{i-1}},$$

то, використовуючи апроксимацію частинної суми гармонічного ряду $H_n = 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n}$ формулою [2] $\ln n + C + \gamma_n$, де $C = 0,577\ldots$ – ейлерова стала, а γ_n – нескінченно мала величина, одержуємо

$$\sum_{k=1}^{2^{i-1}} \frac{1}{2k-1} = \ln 2^i + C + \gamma_{2^i} - \frac{1}{2}\left(\ln 2^{i-1} + C + \gamma_{2^{i-1}}\right)$$

або

$$\sum_{k=1}^{2^{i-1}} \frac{1}{2k-1} = \left(i - \frac{1}{2}(i-1)\right) \ln 2 + \frac{1}{2}C + \gamma_{2^i} - \frac{1}{2}\gamma_{2^{i-1}}.$$

Нехтуючи нескінченно малими величинами, з достатньо високою точністю можемо прийняти

$$\sum_{k=1}^{2^{i-1}} \frac{1}{2k-1} = \frac{1}{2}\left((i+1)\ln 2 + C\right).$$

Отже, математичне сподівання кількості порівнянь

$$E_2 = \frac{1}{2H_N} \sum_{i=1}^{l} \frac{i}{n_i} \left((i+1)\ln 2 + C\right)$$

або

$$E_2 = \frac{1}{2mH_N} \sum_{i=1}^{l} i2^{i-1} \left((i+1)\ln 2 + C\right). \tag{3}$$

Проведемо порівняльний аналіз значення математичного сподівання кількості порівнянь, необхідних для пошуку запису у файлі, у випадку використання методів послідовного перегляду (E_1) та двійкового пошуку (E_2) при розподілі ймовірностей за законом Зіпфа, а також двійкового пошуку у випадку рівномірного розподілу ймовірностей (E) і за законом Зіпфа (E_2). У табл. наведені значення E_1, E_2 і E, обчислені за формулами (2), (3) і (1) відповідно, для різних значень N.

Таблиця 1. Математичне сподівання кількості порівнянь

l	$N = 2^l - 1$	E_1	E_2	E
1	1	1,7331	0,98164	1
2	3	1,79039	1,71669	1,66667
3	7	2,77457	2,54469	2,42857
4	15	4,56614	3,43089	3,26667
5	31	7,72877	4,35333	4,16129
6	63	13,3471	5,29857	5,09524
7	127	23,4266	6,25844	6,05512
8	255	41,6785	7,22797	7,03137
9	511	74,9996	8,20409	8,01761
10	1023	136,264	9,18485	9,00978

На рис.1 зображено графік поведінки математичного сподівання E_1 і E_2 для різних значень N.

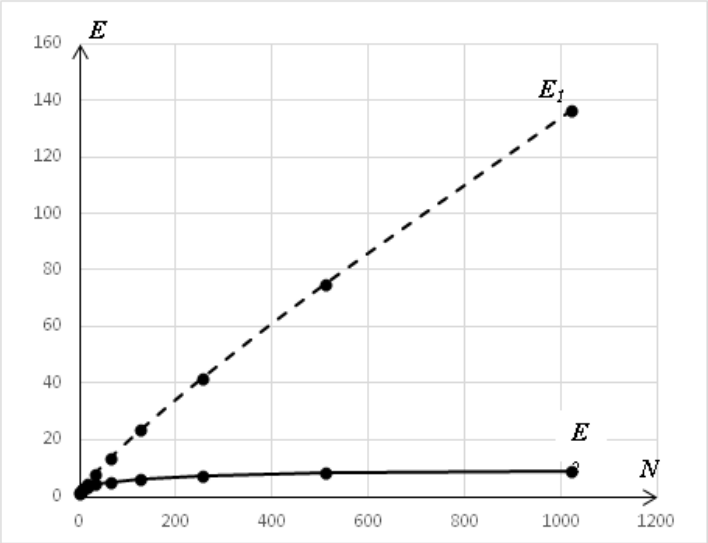

Рис. 1. Графік поведінки математичного сподівання E_1 і E_2 для різних значень N.

Одержані результати показують, що при використанні методу послідовного перегляду для пошуку запису у файлі математичне сподівання зростає дуже швидко зі збільшенням N і є значно більшим, ніж при використанні двійкового пошуку, що свідчить про ефективність останнього. Якщо ж порівнювати метод двійкового пошуку у випадку рівномірного розподілу ймовірностей і розподілу за законом Зіпфа, то у випадку рівномірного розподілу ймовірностей результати є дещо кращими, але ця перевага є незначною.

3. Висновки. У роботі виведено формулу для обчислення математичного сподівання кількості порівнянь, необхідних для пошуку запису у файлі, у випадку розподілу ймовірностей звертання до записів за законом Зіпфа. Зроблено порівняння ефективності методу послідовного перегляду та методу двійкового пошуку у цьому випадку, а також порівняння ефективності двійкового пошуку у випадку рівномірного розподілу ймовірностей і розподілу за законом Зіпфа.

Список використаної літератури

1. Кнут Д. Искусство программирования для ЭВМ. Москва, 2000. Т.3. Сортировка и поиск. 824 с.
2. Мартин Дж. Организация баз данных в вычислительных системах. Москва, 1980. 662 с.
3. Цегелик Г. Г. Моделювання та оптимізація доступу до інформації файлів баз даних для однопроцесорних і багатопроцесорних систем: монографія. Львів, 2010. 192 с.
4. Цегелик Г. Г. Организация и поиск информации в базах данных. Львов, 1987. 176 с.
5. Leipälä T. On the design of one-level indexed sequential files, *Int. J. Comput. Inform.* 1981. Sci. 10, No. 3, P. 177–186.
6. Leiälä T., On optimal multilevel indexed sequential files. *Inform. Process. Lett.* 1982. Vol. 15, No. 5, P. 191–195.

Fundak L. I., Tsehelyk H. H., Hlebena M. I. Effectiveness of the binary search method in database files in the case of a distribution of probabilities of access to records according to the Zipf law.

The main emphasis in solving various tasks using the concept of databases is transferred from the procedures for processing information to the procedures for organizing the stor-

age and retrieval of information in databases. Therefore, the performance of computing systems, focused on processing information in large databases, is mainly determined by the effectiveness of information search methods in database files.

Since most systems of information processing are typical cases of uneven distribution of probabilities of access to file records, the research of the effectiveness of search methods is performed for such standard laws of unequal distribution of probabilities as binary, Zipf's law, generalized law.

The criterion for the effectiveness of the methods is the mathematical expectation of the number of comparisons required to search for a record in a file. You can use various methods to search the record in a file: sequential view; one-level or multi-level block search; binary search; a search method that takes into account the distribution of probabilities of access to records; search methods that use indexes, etc. The effectiveness of these methods for different laws of distribution of likelihood of access to records is different.

We will assume that the database file is organized in ascending order of key values. In the article, a formula is derived for calculating the mathematical expectation of the number of comparisons required to search for a record in a file, in the case of the distribution of the probabilities of accessing records according to the Zipf law. Comparison of the effectiveness of the sequential method and the binary search method in this case is compared, as well as comparison of the efficiency of the binary search in the case of a uniform distribution of probabilities and distribution according to the Zipf law.

The graphs show the dependence of the mathematical expectation of the number of comparisons between the number of records in a file in the case of the distribution of probabilities of accessing records according to the Zipf law.

Keywords: the distribution of the probabilities of access to records with applying to Zipf law, the binary search method, the mathematical expectation.

References

1. Knut, D. (2000). Iskusstvo prohrammirovaniya dlya *ÉVM*. T. 3. Sortirovka y poisk. [The art of programming for computers. Vol. 3. Sort and search]. Moscow [in Russian].
2. Martyn, Dzh. (1980). Orhanizatsiya baz dannykh v vychislytel'nykh sistemakh [Database organization in computing systems]. Moscow [in Russian].
3. Tsehelyk, H. H. (2010). Modelyuvannya ta optymizatsiya dostupu do informatsiyi fayliv baz danykh dlya odnoprotsesornykh i bahatoprotsesornykh system: monohrafiya. [Modeling and optimization of access to information file databases for single-processor and multiprocessor systems]. Lviv [in Ukranian].
4. Tsehelyk, H. H. (1987). Orhanyzatsiya i poisk informatsii v bazakh dannykh. [Organization and search of information in databases]. Lviv [in Russian].
5. Leipälä, T. (1981). On the design of one-level indexed sequential files. *Int. J. Comput. Inform., 10(3)*, 177–186.
6. Leipälä, T. (1982). On optimal multilevel indexed sequential files. *Inform. Process. Lett., 15(5),* 191–195.

Одержано 11.04.2019