UDC 519.68

EFFECTIVENESS OF THE BINARY SEARCH METHOD IN DATABASE FILES IN THE CASE OF A GENERALIZED DISTRIBUTION OF PROBABILITIES OF ACCESS TO RECORDS

L. Fundak, H. Tsehelyk

Ivan Franko National University of Lviv,
Universytetska str., 1, Lviv, 79000,
e-mail: lesya.fundak@lnu.edu.ua, kafmmsep@lnu.edu.ua

Since the main emphasis in solving various problems using the concept of databases is transferred from information processing procedures to procedures for storing and retrieving information in databases, the performance of computer systems focused on processing information in large databases is mainly determined by the efficiency of search methods information in database files. In most systems of information processing there are typical cases of uneven distribution of probabilities of access to records. Among the uneven laws, the most common are the "binary law", the Zipf's law and the generalized law, a partial case of which is a distribution that approximately satisfies the "80-20" rule. For these laws, the mathematical expectation of the number of comparisons required to search for records in a file, in the methods of sequential viewing, one- level, two- level and multi-level block search is calculated. However, in the case of binary search, mathematical expectation is found only for the laws of Zipf and "binary". The work is devoted to the case of the generalized law. The most effective method of binary search in the case of uniform distribution of probabilities of access to records of database files is considered. A formula to calculate the mathematical expectation of the number of comparisons required to find an entry in a file in the case of a generalized law of distribution of the probabilities of access to records is derived. A comparative analysis of the effectiveness of the binary search method in the case of a generalized law of distribution of probabilities of access to records and distribution according to Zipf's law is done. The graphs show the dependence of the mathematical expectation of the number of expectations on the number of records in the file, as well as the results of comparing the effectiveness of methods.

Key words: the generalized distribution of probabilities of access to records, Zipf's law, the binary search method, the mathematical expectation.

1. Introduction

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 storage 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. Some partial results of research of the effectiveness on search methods were obtained by foreign authors. In particular, they reflected in the monographs by D. Knut and J. Martin [1,2]. More complete studies conducted in the works of H. H. Tsehelyk [3,4]. 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.

A formula for calculating the mathematical expectation of the number of comparisons needed to find a record in a file in the case of generalized distribution of the probabilities of accessing records is derived in this article. The effectiveness of the binary search method in the case of generalized distribution of probabilities of access to records and Zipf's law is compared.

2. Theoretical Result

Consider a file that contains N records. Let k_i, $i = 1, 2, .., N$, – the value of the key that characterizes the i-th file entry, and p_i, $i = 1, 2, .., N$, – the probability of accessing to the i-th file entry. We will assume that the database file is sorted in ascending order of key values and the record in file is searched with using binary search method [3]. If the distribution of probabilities of access to records is uniform, that is $p_i = \frac{1}{N}$, $i = 1, 2, .., N$, then this method is most effective. In [3] it is shown that the maximum number of comparisons required to find a record in a file when using this method is

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

and the average number of comparisons is expressed by the formula

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

In the case of non-uniform laws of probability distribution the formula for mathematical expectation of the number of comparisons needed to find a record in a file can only be written if $N = 2^l - 1$, where l – is an integer ($l \geq 2$). This formula looks like this

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

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

Using this formula, in [5] we found an explicit form of mathematical expectation in the case of distribution of probabilities of access to records according to Zipf's law [1]. We also compared the effectiveness of sequential browsing and binary search methods in the case of probability distribution under Zipf's law.

Let's find the mathematical expectation of the number of comparisons required to find a record in a file in the case of a generalized distribution of the probabilities of accessing records [3]. They are calculated by the formula

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

where

$$H_N^{(c)} = 1 + \frac{1}{2^c} + \frac{1}{3^c} + ... + \frac{1}{N^c} = \sum_{k=1}^{N} \frac{1}{k^c}$$

– the partial sum of the generalized harmonic series, $0 < c < 1$.

In this case the mathematical expectation

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

Since

$$\sum_{k=1}^{2^{i-1}} \frac{1}{(2k-1)^c} = 1 + \frac{1}{3^c} + \frac{1}{5^c} + ... + \frac{1}{(2^i-1)^c} =$$

$$= 1 + \frac{1}{2^c} + \frac{1}{3^c} + \frac{1}{4^c} + ... + \frac{1}{(2^i-1)^c} + \frac{1}{(2^i)^c} - \left(\frac{1}{2^c} + \frac{1}{4^c} + \frac{1}{6^c} + ... + \frac{1}{(2^i)^c}\right) =$$

$$= H_{2^i}^{(c)} - \frac{1}{2^c}\left(1 + \frac{1}{2^c} + \frac{1}{3^c} + ... + \frac{1}{(2^{i-1})^c}\right) = H_{2^i}^{(c)} - \frac{1}{2^c} H_{2^{i-1}}^{(c)},$$

then, using the approximation [4]

$$H_n^{(c)} = \frac{1}{1-c} n^{1-c} - C^{(c)} + \gamma_n^{(c)}, \tag{1}$$

where $C^{(c)}$ – some constant, and $\gamma_n^{(c)}$ – an infinitely small value, we obtain,

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

or

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

Neglecting infinitesimally small quantities, with high enough accuracy we can accept that

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

Hence, the mathematical expectation of the number of comparisons will be calculated by the formula

$$E = \frac{1}{H_N^{(c)}} \sum_{i=1}^{l} \frac{i}{n_i^c}\left(\frac{1}{1-c}\, 2^{i(1-c)-1} + \left(\frac{1}{2^c} - 1\right) C^{(c)}\right)$$

or

$$E = \frac{1}{m^c H_N^{(c)}} \sum_{i=1}^{l} i\, 2^{c(i-1)}\left(\frac{1}{1-c}\, 2^{i(1-c)-1} + \left(\frac{1}{2^c} - 1\right) C^{(c)}\right), \tag{2}$$

where in the case $N = 2^l - 1$, we have $n_i = 2^{l-i}$, $m = 2^{l-1}$.

3. Practical Result

We calculate the value of the constant $C^{(c)}$ using formula (1) and examine its dependence on values n and c. The results of the calculations are shown in table1 and Fig.1.

We can see that with increasing n at a given value c the constant $C^{(c)}$ grows very slowly. So the following calculations of the mathematical expectation are almost independent of the value n.

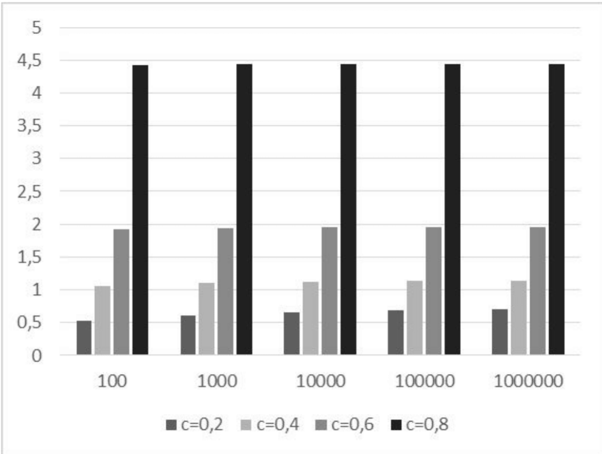

Fig. 1. The constant value $C^{(c)}$depending on n and c

Table 1

The constant value $C^{(c)}$depending on n and c

n	$c = 0,2$	$c = 0,4$	$c = 0,6$	$c = 0,8$
100	0,534934	1,05561	1,92115	4,425
1000	0,608331	1,10325	1,94474	4,43555
10000	0,654677	1,12224	1,95067	4,43722
100000	0,683921	1,1298	1,95216	4,43749
1000000	0,702373	1,13281	1,95254	4,43753

Let us perform a comparative analysis of the value of the mathematical expectation of the number of comparisons required to find a record in a file, in the case of using the binary search method in the generalized distribution of probabilities of access to records. In table 2 it shows the values E calculated by the formula (2) for different values c and value $n = 10000$. Also in this table we can see the mathematical expectation of the

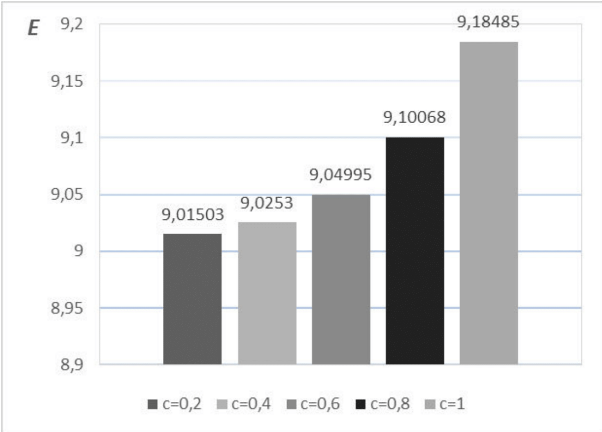

Fig. 2. The mathematical expectation for different values c

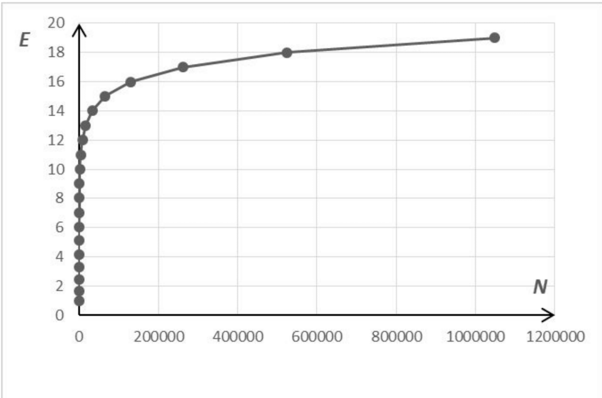

Fig. 3. The mathematical expectation for $c = 0,2$

number of comparisons E_1 and E_2 calculated with the using sequential view method and binary search method in the case of distribution according to Zipf's law.

The Fig. 2 shows the values of the mathematical expectation E at the $l = 10$ ($N = 1023$) for different values c (if $c = 1$we have the Zipf's law). We see that the best result obtained for $c = 0,2$ and in the case of Zipf's law the value the mathematical expectation E will be the greatest. It grows with the growing c.

The Fig. 3 shows the behavior of the mathematical expectation of the number of comparisons needed to find a record in a file, in the case of generalized distribution of the probabilities of accessing records with $c = 0,2$, depending on the number of records N.

4. Conclusions

The paper presents a formula for calculating the mathematical expectation of the number of comparisons required to find a record in a database file in the case of generalized distribution of the probabilities of accessing records. We compare the effectiveness

Table 2

The mathematical expectation of the number of comparisons required to find a record in a file

l	N	E $(c=0,2)$	E $(c=0,4)$	E $(c=0,6)$	E $(c=0,8)$	E_1	E_2
1	1	1,00344	0,991356	0,985677	0,983038	1,7331	0,98164
2	3	1,68083	1,68043	1,68837	1,70122	1,79039	1,71669
3	7	2,4499	2,46091	2,48231	2,51114	2,77457	2,54469
4	15	3,29032	3,30893	3,34006	3,38172	4,56614	3,43089
5	31	4,18352	4,20541	4,24174	4,29202	7,72877	4,35333
6	63	5,11399	5,1358	5,17342	5,22843	13,3471	5,29857
7	127	6,06978	6,08943	6,12554	6,18218	23,4266	6,25844
8	255	7,04221	7,05876	7,09167	7,14767	41,6785	7,22797
9	511	8,02529	8,03856	8,06745	8,12129	74,9996	8,20409
10	1023	9,01503	9,0253	9,04995	9,10068	136,264	9,18485
11	2047	10,0089	10,0166	10,0372	10,0843	249,6	10,169
12	4095	11,0052	11,0109	11,0278	11,071	460,396	11,1557
13	8191	12,0031	12,0072	12,0209	12,0602	854,316	12,1444
14	16383	13,0018	13,0047	13,0157	13,0512	1593,52	13,1346
15	32767	14,0001	14,0031	14,0118	14,0438	2985,83	14,1261
16	65535	15,0006	15,002	15,0089	15,0375	5616,96	15,1187
17	131071	16,0004	16,0014	16,0068	16,0322	10604	16,1127
18	262143	17,0002	17,0009	17,0051	17,0277	20082	17,1093
19	524287	18,0001	18,0006	18,0039	18,0239	38138,9	18,1137
20	1048575	19,0001	19,0004	19,0029	19,0206	72616,3	19,1441

of the binary search method for different values c, as well as in the case of Zipf's law distribution.

First of all, the binary search method in the case of generalized distribution produces much better results than the sequential view method and better than the binary search in the case of Zipf's law.

Also, the best result we obtained for the value $c = 0,2$. The mathematical expectation grows with the growing c. The worst result was in the case of Zipf's law.

We calculate the value of the constant $C^{(c)}$ and examine its dependence on values n and c. With increasing n at a given value c the constant $C^{(c)}$ grows very slowly. That's why the mathematical expectation of the number of comparisons required to find a record in a database file in the case of generalized distribution of the probabilities of accessing records are almost independent of the value n.

We examine the behavior of the mathematical expectation of the number of comparisons needed to find a record in a file.

Calculations show that the use of the binary search method in the case of a generalized law of distribution of the probability of access to records is ineffective. In this case, it would be best to use a method that takes into account the distribution of record access probabilities [6]. This method uses the notion of a conditional average record. In the case of even distribution of access to records, it is the same as the binary search method.

References

1. *Knuth D.* The Art of Computer Programming. Volume 3: Sorting and Searching / D. Knuth. – Moscow, 2000. – 824 p. (In Russian)
2. *Martin J.* Managing the data base environment / J. Martin. – Moscow, 1980. – 662 p. (In Russian).
3. *Tsehelyk H.* Modeling and optimizing of access to database file information for single-processor and multiprocessor systems / H. Tsehelyk. – Lviv, 2010. – 192 p. (In Ukrainian).
4. *Tsehelyk H.* Organization and search of information in databases / H. Tsehelyk. – Lviv, 1987. – 176 p. (In Russian).
5. *Fundak L.* 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. / L. I. Fundak, H. H. Tsehelyk, M. I. Hlebena // Scientific bulletin of Uzhhorod university. Series of mathematics and informatics. – Uzhhorod, 2019. – Issue № 1 (34). – P. 102–107. (In Ukrainian).
6. *Melnychyn A.* The effectiveness of the method of finding information in a database file that takes into account the probability distribution of requesting to records. / A. V. Melnychyn, M. I. Filyak, H. H. Tsehelyk // Visnyk of Vinnytsia Polytechnical Institute. – Vinnytsia, 2006. – Issue № 6. – P. 187–191. (In Ukrainian).

Article: received 08.07.2020
revised 21.09.2020
printing adoption 23.09.2020

ЕФЕКТИВНІСТЬ МЕТОДУ ДВІЙКОВОГО ПОШУКУ ЗАПИСІВ У ФАЙЛАХ БАЗ ДАНИХ У ВИПАДКУ УЗАГАЛЬНЕНОГО РОЗПОДІЛУ ЙМОВІРНОСТЕЙ ЗВЕРТАННЯ ДО ЗАПИСІВ

Л. Фундак, Г. Цегелик

Львівський національний університет імені Івана Франка,
вул. Університетська, 1, Львів, 79000,
e-mail: *lesya.fundak@lnu.edu.ua*, *kafmmsep@lnu.edu.ua*

Оскільки 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, то продуктивність обчислювальних систем, орієнтованих на опрацювання інформації у великих БД, головно визначена ефективністю методів пошуку інформації у файлах баз даних. У більшості систем опрацювання інформації типові, є випадки нерівномірного розподілу ймовірностей звертання до записів. Серед нерівномірних законів найпоширеніший "бінарний закон", закон Зіпфа і узагальнений закон, частковим випадком якого є розподіл, який наближено задовольняє правило "80-20". Для цих законів знайдено математичне сподівання кількості порівнянь, необхідних для пошуку записів у файлі, в методах послідовного перегляду, однорівневого, дворівневого та багаторівневого блокового пошуку. Однак у випадку двійкового пошуку математичне сподівання знайдено тільки для законів Зіпфа та "бінарного". Праця присвячена випадку узагальненого закону. Розглядається найефективніший у випадку рівномірного розподілу ймовірностей звертання до записів файлів баз даних метод двійкового пошуку. Виведено формулу для обчислення математичного сподівання кількості порівнянь, необхідних для пошуку запису у файлі, у випадку узагальненого закону розподілу ймовірностей звертання до записів. Проведено порівняльний аналіз ефективності методу двійкового пошуку у випадку узагальненого закону розподілу ймовірностей звертання до записів і розподілу за законом Зіпфа. На графіках та у таблицях показана залежність математичного сподівання кількості сподівань від кількості записів у файлі, а також результати порівняння ефективності методів.

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