СИСТЕМНИЙ АНАЛІЗ

УДК 004.422.63:519.24 *http://dx.doi.org/10.30970/vam.2026.36.00000*

ПРО ЕФЕКТИВНІСТЬ ОДНОГО МЕТОДУ ПОШУКУ У ВЕЛИКИХ МАСИВАХ ДАНИХ

А. Мельничин, А. Чипурко

Львівський національний університет імені Івана Франка,
вул. Університетська 1, Львів, 79000, Україна,
e-mail: andriy.melnychyn@lnu.edu.ua, andrii.chypurko@lnu.edu.ua

У роботі пропонується та обґрунтовується вдосконалений метод пошуку інформації у великих масивах даних, ключовою особливістю якого є безпосереднє врахування апріорного розподілу ймовірностей звертання до окремих елементів масиву. Актуальність дослідження зумовлена постійним зростанням обсягів неструктурованих даних, де класичні підходи не завжди забезпечують оптимальну швидкість доступу при нерівномірній частоті запитів.

Авторами розроблено алгоритмічну модель, що адаптує структуру пошуку відповідно до статистичних характеристик вхідного потоку запитів. Основна увага приділена дослідженню ефективності запропонованого методу в порівнянні з традиційними інструментами: методом послідовного перегляду та класичним двійковим пошуком. Математичне моделювання та серія чисельних експериментів проводилися для різних сценаріїв розподілу ймовірностей, зокрема для рівномірного розподілу, специфічного "бінарного" закону та емпіричного закону Ціпфа, що описує реальні процеси звернення до інформаційних ресурсів.

Результати порівняльного аналізу демонструють, що для розподілів з високою нерівномірністю (таких як закон Ціпфа) запропонований метод дозволяє суттєво скоротити середню кількість порівнянь, необхідних для знаходження цільового елемента. Встановлено критичні межі ефективності, за яких адаптивний підхід випереджає двійковий пошук. Теоретично доведено, що врахування статистичної природи запитів дозволяє мінімізувати математичне сподівання часу пошуку, що є критично важливим для систем опрацювання великих даних (Big Data). Отримані дані можуть бути використані при проектуванні інтелектуальних систем пошуку та оптимізації структур зберігання даних у сучасних інформаційних комплексах.

Ключові слова: алгоритм пошуку, ефективність пошуку, закони розподілу ймовірностей.

1. Вступ

Алгоритми пошуку є однією з фундаментальних складових інформатики та програмної інженерії. Вони використовуються для знаходження елементів у масивах, структурах даних, базах даних та інформаційних системах. Ефективність пошуку значною мірою визначає продуктивність програмних систем, особливо у випадках роботи з великими обсягами даних.

З розвитком інформаційних технологій та зростанням обсягів даних виникла потреба у вдосконаленні алгоритмів пошуку, що призвело до появи нових методів та оптимізацій класичних підходів. У сучасних дослідженнях велика увага приділяється оптимізації складності алгоритмів [1, 5], використанню ймовірнісних моделей (Баєсівський двійковий пошук) [6], адаптивним методам пошуку (використання прогнозованої ентропії) [7].

У працях [2, 3] досліджена ефективність методів послідовного перегляду і двійкового пошуку інформації у файлах баз даних для різних законів розподілу ймовірностей звертання до записів (рівномірного і "бінарного", закону Ціпфа, узагальненого закону, частковим випадком якого є розподіл, що наближено задовольняє правило "80-20"). За критерій ефективності прийнято середню кількість порівнянь, необхідних для пошуку запису у файлі.

Пропонується застосувати подібні підходи до пошуку у великих масивах даних.

Зважаючи на те, що алгоритми методів послідовного перегляду і двійкового (бінарного) пошуку жодним чином не враховують розподіл ймовірностей звертання до елементів масиву, виникає задача побудови такого методу пошуку, який би суттєво враховував розподіл ймовірностей звертання до даних. З огляду на актуальність підходів до підвищення ефективності саме такий метод пошуку і пропонується в даній роботі.

Проведемо також дослідження ефективності запропонованого методу в порівнянні з методами послідовного перегляду та двійкового пошуку для рівномірного і так названого "бінарного" законів розподілу ймовірностей звертання до даних, а також для закону Ціпфа [2, 3].

2. Постановка задачі та алгоритм методу

Постановка задачі. Нехай маємо довільний достатньо великий масив даних. Позначимо через N – загальну кількість елементів масиву, p_i – ймовірність звертання до i-го елемента масиву, K – значення i – того елемента масиву.

Розглянемо задачу побудови такого метод пошуку, алгоритм якого в повній мірі враховував би розподіл імовірностей звертання до його елементів, та дослідимо ефективність цього методу в порівнянні з ефективністю методів двійкового пошуку та послідовного перегляду для різних відомих законів розподілу ймовірностей звертання до записів.

Алгоритм методу. Подібно до того, як це зроблено у [4], для опису алгоритму методу, введемо поняття середнього за ймовірностями елемента масиву. Вважатимемо, що середнім за ймовірностями серед елементів масиву з порядковими номерами від m до n, де $1 \le m \le n \le N$, буде елемент масиву із порядковим номером s_{avg}, якщо умова

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

досягається для $k = s_{avg}$. Зауважимо, що може бути випадок, коли мінімум досягається для двох різних індексів k. У такому випадку за s_{avg} приймаємо значення меншого із них. Слід зауважити, що відповідні суми можуть бути невизначиними, якщо виконуються певні умови. Тоді ми прийматимемо:

$$\sum_{i=m}^{k-1} p_i = 0, \text{ якщо } k = m, \quad \sum_{j=k+1}^{n} p_j = 0, \text{ якщо } k = m.$$

Припустимо, що у масиві даних потрібно знайти деякий елемент, значення якого становить K. Тоді на першому етапі порівнюватимемо шукане значення K зі значенням елементу масиву з індексом s_{avg}, який є середнім за ймовірностями. За такого порівняння можливі два випадки. Порівняння успішне (шукане значення

та значення, яке є середнім за ймовірностями, співпадають) – успішно завершуємо алгоритм. Якщо співпадіння не відбулося, то, подібно до алгоритму поділу навпіл, [2] із порівняння видно, в якій частині масиву потрібно продовжувати пошук. На наступному кроці K порівнюється зі значенням елемента масиву, який є середнім за ймовірностями у вибраній частині масиву. При успішному порівнянні робота алгоритму закінчується. При неуспішному пошук продовжується у ще меншій частині масиву. Кроки алгоритму продовжуємо, допоки не закінчаться елементи масиву. Таким чином, через скінченну кількість кроків шуканий елемент буде знайдений, якщо він міститься у заданому масиві.

3. Визначення середнього за ймовірностями елементу та оцінка ефективності алгоритму

Розглянемо найбільш вживані випадки розподілу ймовірностей звертання до даних масиву. Припустимо, що ці закони розподілу ймовірностей відомі нам апріорі.

Випадок перший. Вважатимемо, що елементи масиву є розподілені за рівномірним законом розподілу ймовірностей звертання до даних. Тоді алгоритм пошуку в точності співпадатиме з методом двійкового пошуку. У такому випадку середнім за ймовірностями серед елементів з номерами від m до n включно буде елемент з номером s_{avg}, де $s_{avg} = \left[\dfrac{m+n}{2}\right]$.

Випадок другий. Нехай тепер ймовірності звертання до даних розподілені за "бінарним" законом, тобто $p_i = \dfrac{1}{2^i}$, $i = 1, 2, \ldots, N-1$, $p_N = \dfrac{1}{2^{N-1}}$. Тоді для елементів з номерами $2k-1, 2k, 2k+1, \ldots, N$ $(k = 1, 2, \ldots, [N/2])$ середнім за ймовірностями буде елемент з індексом $s_{avg} = 2k$.

Випадок третій. Припустимо, що ймовірності звертання до даних розподілені за законом Ціпфа ($p_i = \dfrac{1}{iH_N}$, $\quad i = 1, 2, \ldots, N$, де $H_N = \sum_{k=1}^{N} \dfrac{1}{k}$ – 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 середнім за ймовірностями серед елементів з номерами від m до n включн, (за умови $n > m+1$), буде елемент з номером $s_{avg} = k$, де k – індекс для якого виконується умова (1). Якщо ж $n = m+1$, то середнім за ймовірностями буде елемент з номером m.

Одержану формулу для знаходження індекса k можна замінити більш простою. 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 $\varepsilon(k)=\varepsilon_1(k)-\varepsilon_2(k)$. З огляду на це, для знаходження k можна використати наближену формулу $k\approx\sqrt{nm+1}$. Оскільки k повинно бути цілим числом, то можемо прийняти

$$k=\left[\sqrt{nm+1}\right].$$

Розглянемо тепер ефективність методу. За критерій ефективності виконання алгоритму тут і надалі буде використовуватися середня кількість порівнянь E, необхідних для пошуку елемента масиву. У випадку рівномірного розподілу ймовірностей звертання до записів середня кількість порівнянь, необхідних для пошуку, виражається формулою

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

де $l=1+[\log_2 N]$.

Якщо ймовірності звертання до елементів відповідатимуть "бінарному" розподілу, то для середньої кількості порівнянь, необхідних для пошуку, одержуємо вираз

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

при $N=2^k$ 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 $N=2^k+1$.

Для знаходження середньої кількості порівнянь, необхідних для пошуку, у разі закону Ціпфа буде використано алгоритмом методу.

4. Порівняльна ефективність алгоритму

У табл. 1 наведено дані практичного розрахунку середньої кількості порівнянь, потрібних для відшукання елемента масиву, для закону Ціпфа, рівномірного та "бінарного" законів розподілу ймовірностей звертання до елементів і деяких N у випадку методу, який суттєво враховує розподіл імовірностей звертання до даних. У табл. 2 та 3 для аналогічних законів розподілу ймовірностей звертання до елементів і для таких же N, приведений розрахунок середньої кількості порівнянь, потрібних для відшукання елемента, у випадку використання методів послідовного перегляду та двійкового пошуку відповідно.

Таблиця 1

Середня кількість порівнянь, необхідних для пошуку елемента, у разі методу, який суттєво враховує розподіл імовірностей звертання до елементів

N	1 023	32 767	1 048 575	33 554 431	1 073 741 823
Рівномірний	9.01	14.00	19.00	24.00	29.00
“Бінарний”	2.00	2.00	2.00	2.00	2.00
Ціпфа	7.999	9.755	12.65	15.55	18.43

Таблиця 2

Середня кількість порівнянь, необхідних для пошуку елемента, у випадку методу послідовного перегляду

N	1 023	32 767	1 048 575	33 554 431	1 073 741 823
Рівномірний	256	16 384	1 048 576	33 554 432	1 073 741 824
“Бінарний”	2.00	2.00	2.00	2.00	2.00
Ціпфа	136.26	2 985.83	72 614.94	1 873 501.47	50 245 287.55

Таблиця 3

Середня кількість порівнянь, необхідних для пошуку елемента, у випадку методу двійкового пошуку

N	1 023	32 767	1 048 575	33 554 431	1 073 741 823
Рівномірний	9.01	14.00	19.00	24.00	29.00
“Бінарний”	417.88	13 372.20	427 910.37	13 693 132.80	438 180 249.62
Ціпфа	362.36	11 362.60	272 702.39	6 981 181.41	186 164 838.50

Зазначимо, що при обчисленні середньої кількості порівнянь, необхідних для пошуку елемента, для різних законів розподілу ймовірностей звертання до записів у випадку методу послідовного перегляду нами використані формули, подібно як в [3]. А для обчислення математичного сподівання у випадку застосування методу двійкового пошуку ми скористались формулою, аналогічною як в [2].

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

яка справджується при $N = 2^l - 1$, де l – будь-яке натуральне число ($l \geq 2$), $n_i = m = n/2^{i-1}$, $m = [N/2] + 1$ (якщо N буде довільне, то в загальному випадку можна використати алгоритмом методу).

5. Висновки

У даному дослідженні запропоновано алгоритм пошуку, який суттєво враховує розподіл ймовірностей звертання до елементів масиву. Для порівняння ефективності

запропонованого алгоритму та методів ослідовного перегляду і двійкового пошуку (для розглянутих законів розподілу ймовірностей звертання до елементів масиву) проведено розрахунок середньої кількості порівнянь, необхідних для пошуку елемента, для деяких N для вхідних масивів даних. Результати порівнянь показали, що побудований метод за ефективністю значно переважає методи послідовного перегляду та двійкового пошуку у випадку розподілу даних за законом Ціпфа. Встановлено, що у випадку "бінарного" розподілу ймовіронстей він співпадає з методом послідовного перегляду, а у випадку рівномірного розподілу ймовірностей – з методом двійкового пошуку.

Список використаної літератури

1. Кормен Т.Г. Вступ до алгоритмів. Переклад з англійської / Т.Г. Кормен, Ч.Е. Лейзерсон, Р.Л. Рівест, Кд. Стайн. – К: К.І.С., 2019. – 1288 с.
2. Мельничин А.В. Ефективність методу двійкового пошуку інформації у файлах баз даних для різних законів розподілу ймовірностей звертання до записів / А.В. Мельничин, Г.Г. Цегелик // Вісн. Львів. ун-ту. Сер. прикл. математика та інформатика. – 2006. – Вип. 11. – С. 213–218.
3. Філяк М. І. Порівняльний аналіз ефективності методу послідовного перегляду для різних законів розподілу ймовірностей звертання до записів / М.І. Філяк, Г.Г. Цегелик, Х.С. Дороцька // Вісн. НУ "Львівська політехніка". Сер. Інформаційні системи та мережі. – 2000. – № 406. – С. 226–231.
4. Цегелик Г.Г. Метод пошуку інформації у файлах баз даних, який враховує розподіл імовірностей звертання до записів / Г.Г. Цегелик, А.В. Мельничин // Фізико-математичне моделювання та інформаційні технології. – 2006. – Вип. 4. – С. 169–177
5. Knuth D.E. The Art of Computer Programming: Volume 3: Sorting and Searching; / D.E. Knuth. – Boston: Addison-Wesley Professional, 1998.
6. Dinitz M. Binary Search with Distributional Predictions / M. Dinitz, Sungjin Im, T. Lavastida, B. Moseley, A. Niaparast, S. Vassilvitskii // Advances in Neural Information Processing Systems (NeurIPS). – 2024. – Vol. 25. – DOI: https://doi.org/10.48550/arXiv.2411.16030.
7. Singh V. Bayesian Binary Search / V. Singh, M. Khanzadeh, V. Davis, H. Rush, E. Rossi, J. Shrader, P. Lio' // Algorithms. – 2025. – Vol. 18, Iss. 8. – Art. 452. – DOI: https://doi.org/10.3390/a18080452.

Стаття: надійшла до редколегії 15.01.2026
доопрацьована 12.02.2026
прийнята до друку 16.03.2026

ON THE EFFICIENCY OF ONE SEARCH METHOD IN LARGE DATA SETS

A. Melnychyn, A. Chypurko

Ivan Franko National University of Lviv,
1, Universytetska str., 79000, Lviv, Ukraine,
e-mail: andriy.melnychyn@lnu.edu.ua, andrii.chypurko@lnu.edu.ua

The paper proposes and justifies an improved method for searching for information in large data sets, the key feature of which is the direct consideration of the a priori probability distribution of accessing individual elements of the array. The relevance of the study is due to the constant growth of unstructured data volumes, where classical approaches do not always provide optimal access speed with uneven frequency of requests.

The authors have developed an algorithmic model that adapts the search structure according to the statistical characteristics of the incoming query stream. The main attention is paid to the study of the effectiveness of the proposed method in comparison with traditional tools: the sequential review method and the classical binary search. Mathematical modeling and a series of numerical experiments were carried out for different probability distribution scenarios, in particular for a uniform distribution, a specific "binary" law and the empirical Zipf law, which describes the real processes of accessing information resources.

The results of the comparative analysis demonstrate that for distributions with high non-uniformity (such as Zipf's law), the proposed method allows to significantly reduce the average number of comparisons required to find the target element. Critical efficiency limits have been established, at which the adaptive approach outperforms binary search. It has been theoretically proven that taking into account the statistical nature of queries allows minimizing the mathematical expectation of search time, which is critically important for Big Data processing systems and highly loaded databases. The obtained data can be used in the design of intelligent search systems and optimization of data storage structures in modern information complexes.

Key words: search algorithm, search efficiency, probability distribution laws.