

doi: 10.32620/oikit.2025.105.16

УДК 004.8

А. Г. Чухрай,
О. Ю. Бенедичук,
О. В. Гавриленко

Метод швидкого пошуку схожих об'єктів

*Національний аерокосмічний університет «Харківський авіаційний інститут»,
Харків, Україна*

Предмет вивчення. У статті запропоновано метод швидкого пошуку найближчих об'єктів у евклідовому просторі, який базується на необхідних умовах близькості, доведених для двох метрик. Метод передбачає випадковий вибір опорних точок, проектування об'єктів у k -вимірний простір та використання спеціальних хеш-структур для прискорення пошуку. Експериментально доведено, що запропонований метод перевершує відомі аналоги за швидкістю виконання. Метою роботи було розроблення ефективного алгоритму пошуку близьких об'єктів, який може застосовуватися в інтелектуальних пошукових системах, технологіях обробки великих даних та системах навчання для кластеризації знань. Завдання. Розробити метод, який би дозволяв швидко знаходити об'єкти, що задовольняють умовам близькості у евклідовому просторі. Довести теоретичні основи методу, включаючи необхідні умови близькості, та експериментально перевірити його ефективність у порівнянні з існуючими підходами. Розглянути два випадки: пошук для впорядкованих m -арних дерев та для рядків з використанням відстані Левенштейна. Використовувані методи. Метод базується на випадковому виборі опорних точок, проектуванні об'єктів у k -вимірний простір та використанні хеш-структур для фільтрації кандидатів. Для оцінки ефективності проведено експерименти з порівнянням часу виконання запропонованого методу з базовими та модифікованими аналогами. Теоретичні основи включають доведення умов близькості, які дозволяють скоротити кількість обчислень. Отримані результати. Експерименти показали, що запропонований метод (Найближчий хеш) перевершує відомі аналоги (Базовий та Модифікований) у 1,67–3,80 рази за швидкістю пошуку, залежно від розміру набору даних. Для дерев використано метрики редагування, які враховують операції перейменування, видалення та вставки вузлів. Для рядків застосовано відстань Левенштейна. Результати підтверджують, що метод ефективно виключає зайві обчислення на першому етапі та прискорює пошук завдяки хеш-матриці. Наукова новизна роботи полягає у введенні необхідних умов близькості, які дозволяють підвищити ефективність пошуку. Запропонований метод може бути використаний у пошукових системах, обробці великих даних та освітніх технологіях. Подальші дослідження будуть спрямовані на оптимізацію вибору опорних точок та розширення застосування методу в практичних задачах.

Ключові слова: швидкий пошук; евклідів простір; необхідні умови близькості; метрика Левенштейна; хеш-структури; кластеризація.

Вступ

Проблема пошуку найближчих об'єктів є однією з ключових у сучасних галузях, пов'язаних із обробкою даних. Вона виникає в задачах кластеризації, рекомендаційних системах, обробці природної мови, біоінформатиці та багатьох інших сферах, де потрібно ефективно знаходити схожі елементи серед великих наборів даних.

Так, наприклад, у машинному навчанні та штучному інтелекті пошук найближчих сусідів використовується для класифікації, аналізу образів і передбачення. У системах розпізнавання зображень або відео необхідно швидко знаходити схожі фрагменти, а в рекомендаційних сервісах — ідентифікувати користувачів або продукти з подібними характеристиками. У обробці текстів та

мовних технологіях актуальною є задача пошуку схожих рядків або документів, що ґрунтується на таких метриках, як відстань Левенштейна. Це важливо для виявлення плагіату, автоматичного виправлення помилок або семантичного аналізу текстів. У біоінформатиці пошук близьких об'єктів застосовується для аналізу послідовностей ДНК, білкових структур та інших біологічних даних. Тут ефективність алгоритмів безпосередньо впливає на швидкість досліджень і точність результатів.

Усі ці галузі стикаються із спільними викликами: необхідністю обробляти великі обсяги даних, зберігати баланс між точністю пошуку та швидкодією, а також адаптувати методи до специфіки конкретних задач. Саме тому розробка універсальних і ефективних алгоритмів пошуку найближчих об'єктів залишається актуальною науковою проблемою.

1. Постановка задачі

В останні десятиліття у сфері штучного інтелекту розроблено та використовується набір методів пошуку "найближчих сусідів" для впорядкування та кластеризації об'єктів [1-3]. Тим не менш, проблема розробки високопродуктивних методів в умовах, коли обчислення відстаней між об'єктами все ще займає певний час пошуку, залишається актуальною. Метод, заснований на необхідних умовах близькості, який є узагальненням умов, доведених раніше для метрики Левенштейна, розроблено, експериментально перевірено та описано в [4, 5].

Загальне формулювання задачі є наступним. Припустимо, що відстань редагування δ між об'єктами певного класу Cl задовольняє умови:

$$\left\{ \begin{array}{l} \delta(X, Y) \geq 0; \\ \delta(X, X) = 0; \\ \delta(X, Y) = \delta(Y, X); \\ \delta(X, Z) \leq \delta(X, Y) + \delta(Y, Z). \end{array} \right. \quad (1)$$

Певний об'єкт rt класу Cl та набір об'єктів $ET = (et_1, et_2, \dots, et_n)$ одного класу. Потрібно знайти всі et_i з набору ET , такі що відстань δ між et_i і rt не більше за задане додатне ціле число λ . Формально потрібно знайти множину $ET_s = \{et_{s1}, et_{s2}, \dots, et_{sl}\}$, таку що $\forall et_{si} \in ET_s \subseteq ET : \delta(et_{si}, rt) \leq \lambda, \lambda \in N, l \leq n$.

Запропонований метод складається з двох кроків.

1-й крок. k елементів $o_1, o_2, \dots, o_k, k \leq n$ випадковим чином вибираються з множини ET . Ці елементи розглядаються як k обертається в k -вимірному евклідовому просторі E^k . Після цього кожен елемент et_i з набору ET пов'язаний з E^k координати точки, які дорівнюють відстаням до опорних точок, тобто $P(et_i)_j = \delta(et_i, o_j), i = \overline{1, n}, j = \overline{1, k}$.

2-й крок. Об'єкт rt пов'язаний у просторі E^k з координатами точки $P(rt)_j = \delta(rt, o_j), j = \overline{1, k}$. Відстані обчислюються лише між об'єктом rt і ті об'єкти, відповідні точки яких у просторі E^k розташовані найближче до точки

$P(rt)$.

Для визначення близькості точок у евклідовому просторі необхідно ввести умови близькості для заданих об'єктів X , Y та Z класу Cl .

Твердження 1. Для заданого об'єкта X , Y та Z класу Cl відстань δ між ними задовольняє умови (1) та наступну нерівність $\forall X, Y, Z \quad \delta(X, Y) \geq |\delta(X, Z) - \delta(Z, Y)|$ є правдою.

Доказ. Розглянемо дві нерівності трикутника: $\delta(X, Z) \leq \delta(X, Y) + \delta(Y, Z)$; $\delta(Y, Z) \leq \delta(Y, X) + \delta(X, Z)$. З першої нерівності випливає $\delta(X, Z) - \delta(Y, Z) \leq \delta(X, Y)$, з другого – $\delta(Y, Z) - \delta(X, Z) \leq \delta(Y, X)$. Об'єднавши обидва вирази та використовуючи властивість симетрії, отримуємо систему нерівностей:

$$\begin{cases} \delta(X, Y) \geq \delta(X, Z) - \delta(Z, Y); \\ \delta(X, Y) \geq \delta(Z, Y) - \delta(X, Z), \end{cases}$$

яку можна переписати як: $\delta(X, Y) \geq |\delta(X, Z) - \delta(Z, Y)|$, що і треба доказати.

Твердження 2. Для заданих об'єктів et_i і et_j класу Cl , відстань δ між якими задовольняє умови (1) та не перевищує порогового значення λ , відповідні точки $P(et_i)$ і $P(et_j)$ простору E^k розташовані на відстані не більше ніж $\lambda\sqrt{k}$, тобто $\forall i \forall j \neq i \quad \delta(et_i, et_j) \leq \lambda : \rho(P(et_i), P(et_j)) \leq \lambda\sqrt{k}$.

Доказ. З визначення метрики E^k наступним чином: $\rho(P(et_i), P(et_j)) = \sqrt{(P(et_i)_1 - P(et_j)_1)^2 + (P(et_i)_2 - P(et_j)_2)^2 + \dots + (P(et_i)_k - P(et_j)_k)^2}$. Згідно з пропозицією 1: $|\delta(et_i, o_1) - \delta(et_j, o_1)| \leq \delta(et_i, et_j) \dots, |\delta(et_i, o_k) - \delta(et_j, o_k)| \leq \delta(et_i, et_j)$. Отже, транзитивно: $|P(et_i)_1 - P(et_j)_1| \leq \lambda \dots, |P(et_i)_k - P(et_j)_k| \leq \lambda$ і, отже: $\sqrt{(P(et_i)_1 - P(et_j)_1)^2 + \dots + (P(et_i)_k - P(et_j)_k)^2} \leq \sqrt{\lambda^2 k} = \lambda\sqrt{k}$, що і треба доказати.

Твердження 3. Для заданих об'єктів et_i і et_j класу Cl відстань δ між якими задовольняє умови (1) та не перевищує порогового значення λ , точка $P(et_j)$ розташована у просторі E^k в гіперкубі з центром у $P(et_i)$ зі стороною, що дорівнює 2λ .

Доказ. Згідно з твердженням 1, можна отримати такі системи нерівностей:

$$\left\{ \begin{array}{l} |\delta(et_i, o_1) - \delta(et_j, o_1)| \leq \delta(et_i, et_j); \\ |\delta(et_i, o_2) - \delta(et_j, o_2)| \leq \delta(et_i, et_j); \\ \dots \\ |\delta(et_i, o_k) - \delta(et_j, o_k)| \leq \delta(et_i, et_j), \end{array} \right. \Rightarrow \left\{ \begin{array}{l} P(et_j)_1 \geq P(et_i)_1 - \lambda; \\ P(et_j)_1 \leq P(et_i)_1 + \lambda; \\ P(et_j)_2 \geq P(et_i)_2 - \lambda; \\ P(et_j)_2 \leq P(et_i)_2 + \lambda; \\ \dots \\ P(et_j)_k \geq P(et_i)_k - \lambda; \\ P(et_j)_k \leq P(et_i)_k + \lambda. \end{array} \right. \quad (2)$$

Геометричний сенс системи нерівностей (2) – це гіперкуб з центром у $P(et_i) = (\delta(et_i, o_1), \delta(et_i, o_2), \dots, \delta(et_i, o_k))$ з довжиною сторони 2λ , що і треба доказати.

Твердження 4. Для заданих об'єктів et_i і et_j класу Cl , відстань δ між якими задовольняє умови (1) та не перевищує порогового значення λ , абсолютне значення різниці відстаней від точок $P(et_i)$ і $P(et_j)$ до початку координат у просторі E^k не перевищує $\lambda\sqrt{k}$, тобто $|\rho(P(et_i), 0) - \rho(P(et_j), 0)| \leq \lambda\sqrt{k}$.

Доказ. Згідно з властивістю евклідового метричного простору (нерівність трикутників) $|\rho(P(et_i), 0) - \rho(P(et_j), 0)| \leq \rho(P(et_i), P(et_j))$. З іншого боку, згідно з твердженням 2: $\rho(P(et_i), P(et_j)) \leq \lambda\sqrt{k}$. З вищесказаного можна отримати: $|\rho(P(et_i), 0) - \rho(P(et_j), 0)| \leq \rho(P(et_i), P(et_j)) \leq \lambda\sqrt{k}$, отже $|\rho(P(et_i), 0) - \rho(P(et_j), 0)| \leq \lambda\sqrt{k}$, що і треба доказати.

Твердження 5. Припустимо, що $u, w \in R$ і $u, w > 0$. Тоді з $[u] \leq w$ слідує: $[u] \leq [w]$, де $[u], [w]$ є цілими частинами чисел u і w відповідно.

Доказ. Є два випадки, коли $[u] \leq w$ істинно: $[u] = [w]$ і $[u] < [w]$. Узагальнюючи обидва, ми можемо отримати умову $[u] \leq [w]$. Також очевидно, що припущення $[u] > [w]$ не могло бути істинним для $[u] \leq w$, отже те, що потрібно було довести $[u] \leq [w]$ є істинним.

Твердження 6. Припустимо, що $u, v, w \in R$ і $u, v, w > 0$. Тоді з $|u - v| \leq w$ слідує: $[u] - [v] \leq [w] + 1$, де $[u], [v], [w]$ – цілі частини чисел u, v, w відповідно.

Доказ. Спочатку розглянемо випадок, коли $u \geq v$. Тоді з $u \leq w + v$ можна отримати $[u] \leq w + v$, враховуючи $u \geq [u]$, крім того $[u] \leq w + [v] + 1$, враховуючи $v < [v] + 1$. Далі, згідно з твердженням 5 та тим фактом, що будь-яке додатне число більше за будь-яке від'ємне, якщо $[u] - [v] \leq w + 1$, тоді $[u] - [v] \leq [w] + 1$. Другий випадок $v > u$ можна розглянути таким самим чином і отримати наступну нерівність: $[v] - [u] \leq [w] + 1$. Підсумовуючи обидва випадки: $[u] - [v] \leq [w] + 1$, що і треба доказати.

Визначення 1. Розмір об'єкта et_i класу Cl – кількість його елементів, позначена як $\overline{et_i}$. Наприклад, якщо et_i є рядком "home", тоді $\overline{et_i} = 4$ (довжина рядка); якщо et_i – це дерево, зображене на рис. 1, тоді $\overline{et_i} = 7$ (кількість вершин).

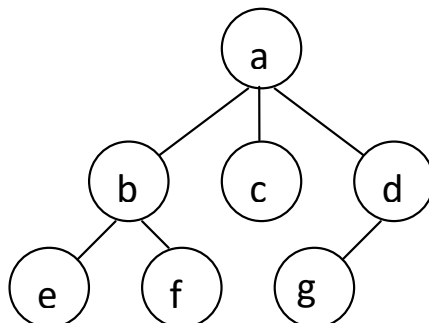


Рис. 1. Приклад дерева як об'єкта

Твердження 7. Якщо абсолютне значення різниці між розмірами об'єктів et_i і et_j з класу Cl , відстань δ між якими задовольняє умови (1), більше ніж λ , тоді відстань між цими об'єктами також більша за λ , тобто $(\overline{et_i} - \overline{et_j}) > \lambda \Rightarrow (\delta(et_i, et_j) > \lambda)$.

Доказ. Це твердження є очевидним і випливає з того факту, що якщо розміри об'єктів відрізняються на λ , то щоб перетворити об'єкт et_i в об'єкт et_j або навпаки, має бути виконано щонайменш λ видалень елементів у найкращому випадку та більше λ – в інших випадках.

Після того, як необхідні умови близькості об'єктів будуть доведені, суть запропонованого методу можна буде описати детальніше.

На першому кроці, після вибору випадкового k з набору ET та розрахунку координат точок $P(et_i)$, відстані від точок $P(et_i)$ до початку координат у просторі E^k можна було б отримати як: $\rho(P(et_i), 0) = \sqrt{P(et_i)_1^2 + P(et_i)_2^2 + \dots + P(et_i)_k^2}$.

Далі формуємо матрицю D , що є розподілом відстаней у просторі E^k для точок $P(et_i)$ до початку координат. Для цього вводиться множина $\Psi = \{[\rho(P(et_1), 0)], [\rho(P(et_2), 0)], \dots, [\rho(P(et_n), 0)]\} = \{\psi_1, \psi_2, \dots, \psi_z\}, z \leq n$, де $[\rho(P(et_i), 0)]$ означає цілу частину $\rho(P(et_i), 0)$.

Кожен індекс $\psi_i \in \Psi^\xi$ пов'язаний з набором цілих чисел $IND_i = \{ind_{i1}, ind_{i2}, \dots, ind_{iw}\}$, які є індексами об'єктів з відстанню від початку координат, що дорівнює ψ_i . У цьому випадку умова $\forall q \in \{1, \dots, w\}, ind_{iq} \in \{1, \dots, n\}, \exists [\rho(P(et_{ind_{iq}}), 0)] = \psi_i$ істинна. Розглянута матриця D має розмір $(\max(\Psi) - \min(\Psi) + 1) \times (\max\{|IND_1|, \dots, |IND_z|\})$. Використання допоміжної множини IND_i , призначено: $D_{\psi_i q} = ind_{iq}$ що означає,

що рядок ψ_i матриці D містить індекси об'єктів, для яких ціла частина відстані в просторі E^k до початку координат дорівнює ψ_i .

2. Для цільового об'єкта rt потрібно знайти рядок з індексом $[\rho(P(rt), 0)]$ у матриці D . Після цього, згідно з твердженнями 4 та 6, необхідно переглянути сусідні рядки матриці D з індексами з множини $\Psi_1 = \{[\rho(P(rt), 0)] - [\lambda\sqrt{k}] - 1, [\rho(P(rt), 0)] - [\lambda\sqrt{k}], \dots, [\rho(P(rt), 0)] - 1, [\rho(P(rt), 0)], [\rho(P(rt), 0)] + 1, \dots, [\rho(P(rt), 0)] + [\lambda\sqrt{k}], [\rho(P(rt), 0)] + [\lambda\sqrt{k}] + 1\} = \{\psi_{1_1}, \psi_{1_2}, \dots, \psi_{1_v}\}$, та $\Psi_1 \subset \Psi$, $v \leq z$.

Тоді, переглядаючи рядки елементів матриці D з індексом ψ_{1_t} , тобто $D_{\psi_{1_{tq}}}$, подальший відбір "кандидатів" можна знайти серед найближчих об'єктів: по-перше, шляхом перевірки умов з твердження 7: $|\overline{et_i} - \overline{et_j}| > \lambda$, і по-друге, перевіряючи умову твердження 3: чи $P(et_{D_{\psi_{1_{tq}}}})$ лежить у гіперкубі з центром у точці $P(rt)$ і стороною 2λ . Зрештою, якщо $P(et_{D_{\psi_{1_{tq}}}})$ знаходиться в межах гіперкуба, тоді розраховується відстань $\delta(rt, et_{D_{\psi_{1_{tq}}}})$ між об'єктами rt і $et_{D_{\psi_{1_{tq}}}}$.

2. Опис методу

Розглядаються два випадки: 1) Об'єкти класу Cl є впорядкованими m -арними деревами, $m \in N$; 2) об'єкти класу Cl є рядками.

1-й випадок. Об'єкти класу Cl є впорядкованими m -арними деревами. Тоді формулювання задачі таке. Задано дерево rt і набір дерев $ET = (et_1, et_2, \dots, et_n)$, $i = \overline{1, n}$. Потрібно знайти всі дерева et_i , такі що відстань $\delta(rt, et_i)$ не більше за задане додатне ціле число λ .

У цьому випадку один із показників для δ можна вибрати. Метрику, використану в [7], та метрику з [4, 5] можна розглядати як можливі альтернативи. Різниця між цими двома метриками полягає в наборі допустимих операцій редагування дерева: кожна метрика дозволяє операції перейменування, видалення та вставки вузлів дерева, але перша розглядає дві останні операції як застосовні лише до листя дерева, тобто до вершин без нащадків, тоді як друга застосовує їх до будь-якого вузла дерева.

Далі буде детально розглянуто суть вставки та видалення вузлів дерева в другій метриці. В результаті операції вставки деякі або всі нащадки батьківського вузла для вставленого вузла перетворюються на нащадків вставленого вузла. Після операції видалення всі нащадки видаленого вузла перетворюються на нащадків його батьківського вузла. Розглянемо конкретні приклади.

Приклад 1. Два впорядковані бінарні дерева X та Y представлено на рис. 2. Відстань редагування між X та Y у метриці Селькова [7] дорівнює 6.

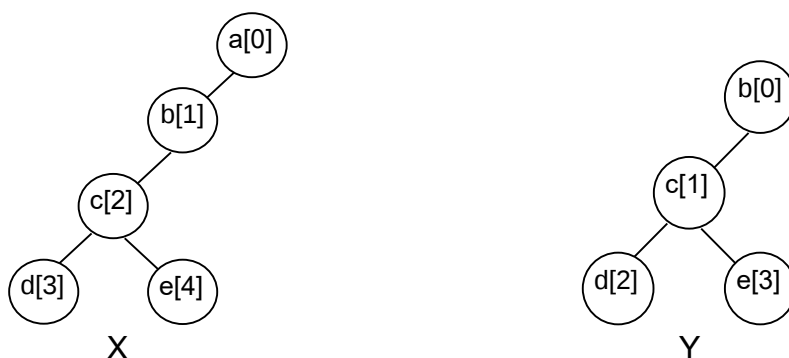


Рис. 2. Впорядковані бінарні дерева

Мінімальний набір операцій редагування, які перетворюють X в Y, включає операції:

- 1) замінити вузол «а» з індексом 0 на назву «b»;
- 2) замінити вузол «b» з індексом 1 та назвою «с»;
- 3) видалити кінцевий вузол «е» з індексом 4;
- 4) видалити кінцевий вузол «d» з індексом 3;
- 5) замінити назву вузла «с» з індексом 2 на назву «d»;
- 6) вставити правий дочірній елемент «е» до вузла «b» з індексом 1;

Для метрики, що використовується в [7, 8], відстань редагування між X та Y буде дорівнювати 1, оскільки для перетворення Y в X необхідно видалити вузол "a" з індексом 0.

Приклад 2. Два впорядковані потрійні дерева X1 та Y1 зображено на рис.3.

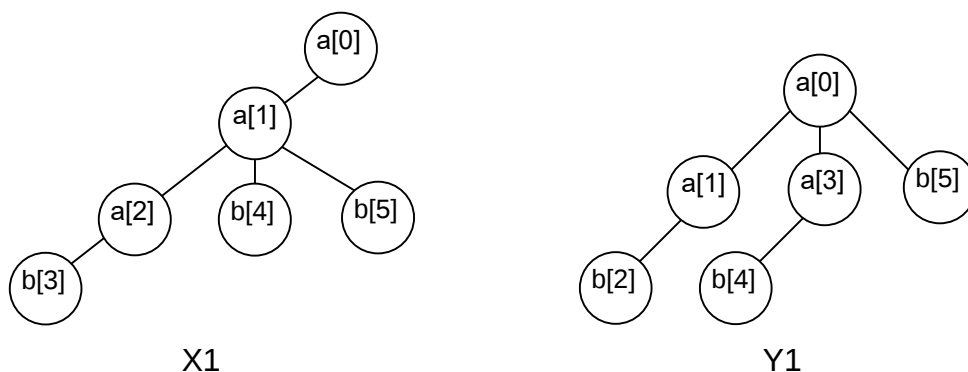


Рис. 3. Впорядковані тернарні дерева

Згідно з метрикою Селькова [6] відстань редагування між X1 та Y1 дорівнює 7. Мінімальний набір операцій редагування, які перетворюють X в Y, включає операції:

- 1) вставити дочірній елемент «а» до вузла «а» з індексом 0;
- 2) вставити дочірній елемент «b» до вузла «а» з індексом 0;
- 3) вставити дочірній елемент «b» до щойно вставленого вузла «а»;
- 4) видалити листок «b» з індексом 3;
- 5) видалити листок «b» з індексом 4;
- 6) видалити листок «b» з індексом 5;
- 7) замінити назву вузла «а» з індексом 2 на назву «b».

Для другої метрики відстань між X1 та Y1 дорівнює 3, оскільки для

перетворення $X1$ в $Y1$ необхідно виконати операції:

- 1) видалити вузол «а» з індексом 0;
- 2) вставити дочірній елемент «b» до вузла «b» з індексом 4;
- 3) замінити назву вузла «b» з індексом 4 на назву «а».

Як видно з наведених вище прикладів, дерева є більш близькими, якщо використовується друга метрика. Загалом, критерії вибору метрики слід визначати залежно від конкретної практичної задачі.

2-й випадок. Об'єкти класу CI є рядками. У цьому випадку використання відстані Левенштейна, яка є мінімальною кількістю операцій редагування рядка для його перетворення в інший, є найкращим вибором. Це було доведено теоретично та емпірично в [4, 5], де розглядається задача пошуку подібних рядків.

3. Експериментальні дослідження методу

Експериментальні дослідження цього методу було проведено для метрики Чжана та Шаші [8]. Для порівняння результатів було обрано метод, описаний у [1], де також було додано умову, сформульовану у твердженні 7, щоб виключити серію додаткових "дорогих" розрахунків відстані редагування дерев.

Для випадково згенерованих дерев у кожному методі час, витрачений на розрахунок відстані редагування, було виключено із загального часу другого етапу пошуку. Результати експериментального дослідження множин з 10, 100 та 1000 дерев показано на рис. 4-6.

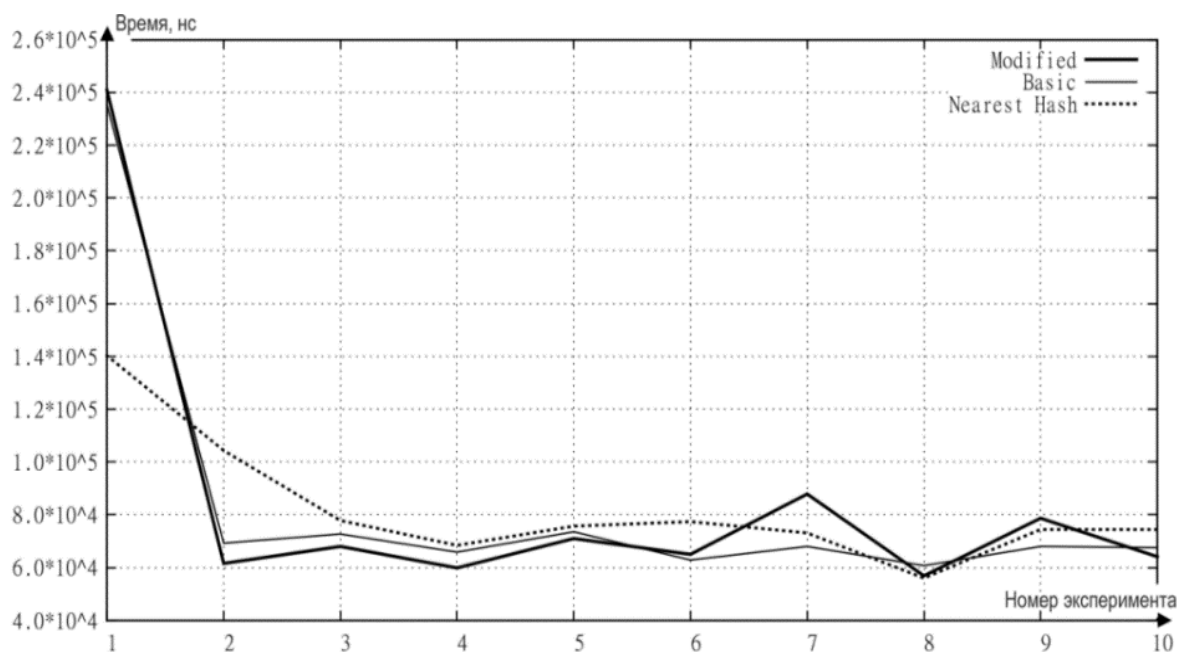


Рис. 4. Результати експериментів для списку з 10 дерев

В експериментах, показаних на рис. 4-5, цільовими були перші 10 дерев з початкового списку.

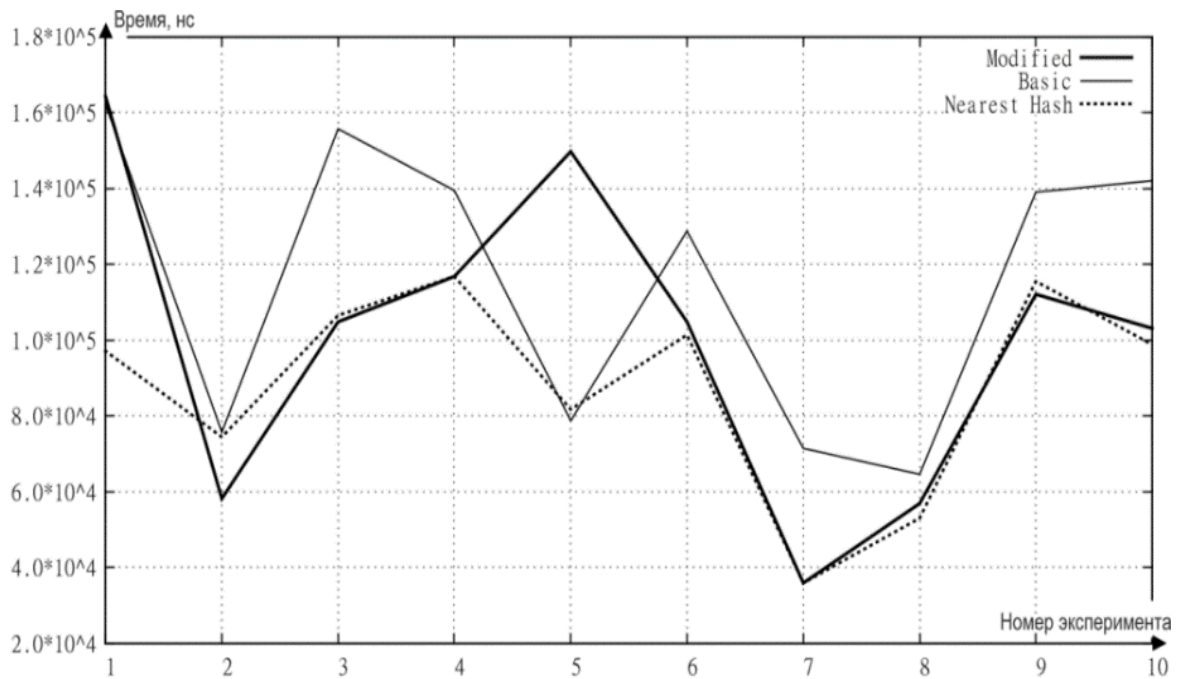


Рис. 5.Результати експерименту для списку зі 100 дерев

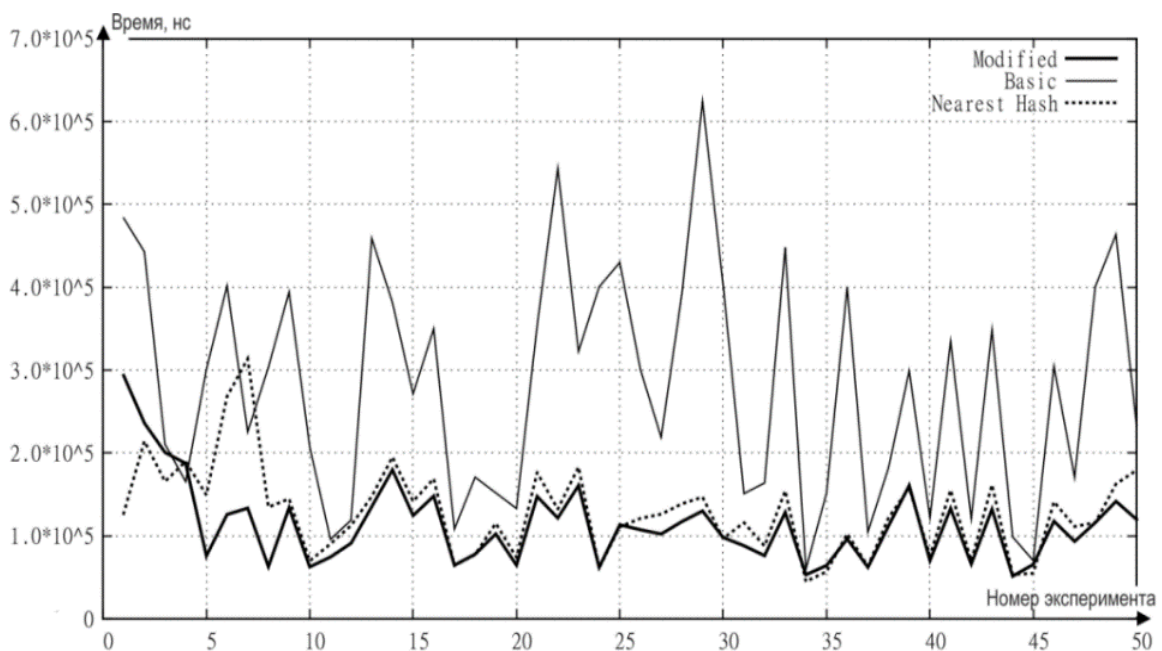


Рис. 6.Результати експерименту для списку з 1000 дерев

В експерименті на рис. 6 цілцю пошуку були перші 50 дерев, які позначені на горизонтальній осі. Як видно з рисунків, запропонований метод пошуку (Найближчий хеш) перевершує відомі (Базовий та Модифікований) за продуктивністю в 1,71 та 1,67 рази – у першому експерименті, в 1,66 та 1,69 рази – у другому експерименті та в 2,30 та 3,80 рази – у третьому експерименті. Таким чином, доведено емпірично теоретичне збільшення продуктивності завдяки виключенню зайвих об'єктів на першому кроці та побудові хеш-матриці на другому.

3. Висновок

Представлено вдосконалений метод ймовірнісного пошуку об'єктів близькості з введенням необхідних теоретично обґрунтованих умов близькості об'єктів, що дозволяє покращити ефективність пошуку. Запропонований метод може бути використаний для швидкого та точного пошуку в інтелектуальних пошукових системах, технології великих даних, а також в інтелектуальних системах навчання для вирішення проблем кластеризації знань та аналізу відповідей учнів для ефективного педагогічного зворотного зв'язку.

Подальші дослідження будуть зосереджені на нових експериментальних дослідженнях, отриманих за допомогою цього методу: порівняння з іншими відомими методами, вибір кількості опорних точок, раціональний вибір конкретних опорних точок, а також різні застосування методу в практичних завданнях.

Список літератури

1. Бустос, Б., Наварро, Г., Чавес, Е.: Методи вибору опорних точок для пошуку близькості в метричних просторах. *Pattern Recognition Letters* 24(14), 2357-2366 (2003).
2. Батько, М., Фальчі, Ф., Луккеше, К. та інші: Створення системи пошуку схожості зображень у веб-масштабі. *Мультимедійні засоби та програми* 3(47), 599–629 (2010).
3. Зезула, П., Амато, Г., Догнал, В., Батько, М.: Пошук подібності: підхід метричного простору. Springer, Нью-Йорк (2006).
4. Чухрай, А.Г.: Метод швидкого пошуку "подібних" реляційних кортежів, відношень. *Радіоелектронні та комп'ютерні системи* 2 (2), 64 – 69 (2003)
5. Кулік, А., Чухрай, А., Завгородній, А.: Методи виявлення подібних рядків. У: *Матеріали Східно-Західного нечіткого колоквиуму*, с. 38 – 47. IPM, Циттау, Німеччина (2005)
6. Селков, С.М.: Проблема редагування від дерева до дерева. *Information Processing Letters* 6(6), 184–186 (1977).
7. Тай, К.К.: Задача корекції від дерева до дерева. *Журнал ACM* 26(3), 422 – 433 (1979).
8. Чжан, К., Шаша, Д.: Прості швидкі алгоритми для визначення відстані редагування між деревами та пов'язані з ними проблеми. *Товариство промислової та прикладної математики. Журнал з обчислювальної техніки* 18(6), 1245 – 1262 (1989).

References

1. Bustos, B., Navarro, G., Chavez, E.: Pivot selection techniques for proximity searching in metric spaces. *Pattern Recognition Letters* 24(14), 2357-2366 (2003).
2. Batko, M., Falchi, F., Lucchese, C. and others: Building a web-scale image similarity search system. *Multimedia Tools and Applications* 3(47), 599 – 629 (2010).
3. Zezula, P., Amato, G., Dohnal, V., Batko, M.: Similarity search: the metric space approach. Springer, New York (2006).
4. Chukhray, A. G.: Quick Search Method "similar" relational tuples relations *Radioelectronic and Computer Systems* 2 (2), 64 – 69 (2003)
5. Kulik, A., Chukhray, A., Zavgorodniy, A.: Similar strings detecting methods. In: *Proceedings of the East-West Fuzzy Colloquium*, pp. 38 – 47. IPM, Zittau, Germany

(2005)

6. Selkow, S. M. : The tree-to-tree editing problem. Information Processing Letters 6(6), 184–186 (1977).

7. Tai, K. C.: The tree-to-tree correction problem. Journal of the ACM 26(3), 422 – 433 (1979).

8. Zhang, K., Shasha, D.: Simple fast algorithms for the editing distance between trees and related problems. Society for Industrial and Applied Mathematics Journal on Computing 18(6), 1245 – 1262 (1989).

Надійшла до редакції 22.08.2025, розглянута на редколегії 22.08.00.2025

Fast search for nearby objects

The subject of study. The article proposes a method of fast search for the closest objects in Euclidean space, which is based on the necessary proximity conditions proved for two metrics. The method involves randomly selecting landmarks, projecting objects into a k-dimensional space, and using special hash structures to speed up the search. It has been experimentally proven that the proposed method outperforms known analogs in terms of execution speed. The aim of the work was to develop an efficient algorithm for finding close objects that can be used in intelligent search engines, big data processing technologies, and learning systems for knowledge clustering. Tasks. To develop a method that would allow to quickly find objects that satisfy the proximity conditions in Euclidean space. Prove the theoretical foundations of the method, including the necessary proximity conditions, and experimentally test its effectiveness in comparison with existing approaches. Consider two cases: search for ordered m-ary trees and for strings using the Levenshtein distance. Methods used. The method is based on randomly selecting landmarks, projecting objects into a k-dimensional space, and using hash structures to filter candidates. To evaluate the effectiveness, experiments have been conducted comparing the execution time of the proposed method with the basic and modified analogs. The theoretical foundations include the proof of proximity conditions that reduce the number of computations. Results. Experiments have shown that the proposed method (Nearest Hash) outperforms the known analogs (Basic and Modified) by 1.67-3.80 times in terms of search speed, depending on the size of the dataset. For trees, we use editing metrics that take into account the operations of renaming, deleting, and inserting nodes. For strings, the Levenshtein distance is used. The results confirm that the method effectively eliminates unnecessary computations at the first stage and speeds up the search due to the hash matrix. The scientific novelty of the work lies in the introduction of the necessary proximity conditions that allow to increase the search efficiency. The proposed method can be used in search engines, big data processing, and educational technologies. Further research will be aimed at optimizing the choice of landmarks and expanding the application of the method in practical problems.

Key words: fast search; Euclidean space; necessary proximity conditions; Levenshtein metric; hash structures; clustering.

Відомості про авторів:

Чухрай Андрій Григорович – професор, доктор технічних наук, професор кафедри інженерії програмного забезпечення Національного аерокосмічного

університету «Харківський авіаційний інститут», Харків, Україна,
ел. пошта: a.chukhray@khai.edu , ORCID 0000-0002-8075-3664.

Бенедичук Олександр Юрійович – аспірант кафедри інженерії програмного забезпечення Національного аерокосмічного університету «Харківський авіаційний інститут», Харків, Україна,
ел. пошта: o.y.benedychuk@khai.edu, ORCID 0009-0001-2225-1973.

Гавриленко Олена Володимирівна – доцент, канд. техн. наук, доцент кафедри систем управління літальних апаратів Національного аерокосмічного університету «Харківський авіаційний інститут», Харків, Україна,
ел. пошта: o.havrylenko@khai.edu, ORCID 0000-0001-5227-9742.

About the authors:

Chukhray Andriy – professor, doctor of technical sciences, professor of the department of software engineering of the National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine, e-mail: a.chukhray@khai.edu, ORCID 0000-0002-8075-3664.

Benedychuk Oleksandr – PhD student of the Department of Software Engineering of the National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine, e-mail: o.y.benedychuk@khai.edu, ORCID 0009-0001-2225-1973.

Havrylenko Olena – Associate Professor, Candidate of Technical Sciences, Associate Professor of the Department of Aircraft Control' Systems of the National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine, e-mail: o.havrylenko@khai.edu, ORCID 0000-0001-5227-9742.