Методика шифрування даних на основі технології ідентифікації блоків матричним способом

Олександр Левкуша[1], Юрій Пепа[2]

[1,2] Державний університет інформаційно-комунікаційних технологій
вул. Солом'янська, 7, м. Київ, Україна, 03037
[1] levkusha_alexandr@ukr.net, https://orcid.org/0009-0001-6064-0822
[2]yurka14@ukr.net, https://orcid.org/0000-0003-2073-1364

Received 03.02.2025, accepted 29.03.2025
https://doi.org/10.32347/st.2025.3.1208

Анотація. У статті запропоновано матричний метод ідентифікації блоків даних на серверних платформах, які працюють із великими масивами інформації. Використання лінійної алгебри, зокрема сингулярного розкладу (SVD), дозволяє ефективно аналізувати структуру даних, зменшувати надлишковість та оптимізувати процес кластеризації. Запропонований метод включає попередню нормалізацію даних, обчислення косинусної міри подібності та адаптивний алгоритм обробки, що забезпечує високий рівень точності ідентифікації навіть при модифікації вхідних блоків. Досліджено обчислювальну складність алгоритму, показано переваги використання методів зменшення розмірності та оптимізації ресурсів. Проведені експериментальні дослідження довели ефективність методу в умовах високої варіативності даних, що дозволяє його застосовувати у кібербезпеці, хмарних обчисленнях і розподілених системах. Запропонований підхід зменшує час обробки запитів та підвищує точність ідентифікації, що критично важливо для моніторингу інформаційних потоків та забезпечення цілісності даних. Отримані результати підтверджують доцільність впровадження цього методу в інформаційних системах для підвищення продуктивності обробки великих даних.

Ключові слова: матричний метод, сингулярний розклад, ідентифікація блоків даних, оптимізація обчислень, кластеризація, великі дані, кібербезпека.

Олександр Левкуша
аспірант кафедри технічних систем кіберзахисту

Юрій Пепа
к.т.н., доцент
професор кафедри технічних систем кіберзахисту

ВСТУП

Розвиток інформаційних технологій супроводжується постійним зростанням обсягів даних, що зберігаються та обробляються на серверних платформах.

Одним із важливих аспектів управління інформаційними потоками є процес ідентифікації блоків даних, який дозволяє визначати їх приналежність, структуру та взаємозв'язки між окремими фрагментами інформації. Оскільки дані на сервері можуть змінюватися, оновлюватися та дублюватися, їх ефективна ідентифікація стає критичною задачею, від вирішення якої залежить швидкість обробки запитів, рівень навантаження на обчислювальні ресурси та загальна продуктивність системи [1 - 4].

Одним із перспективних підходів до вирішення цієї проблеми є використання матричних методів, які дозволяють

здійснювати ідентифікацію блоків даних на основі математичних перетворень та алгоритмів лінійної алгебри. Застосування матричних операцій у цьому контексті сприяє оптимізації процесів обробки інформації, підвищенню точності та швидкості аналізу взаємозв'язків між даними, а також зменшенню обчислювальних витрат [5 - 8].

У представленому дослідженні розглядається розробка та впровадження матричного підходу до ідентифікації блоків даних, що дозволяє систематизувати та автоматизувати процес визначення інформаційних об'єктів у великих масивах даних. Особлива увага приділяється математичному моделюванню, розробці алгоритму, аналізу його обчислювальної складності, а також практичній оцінці ефективності на тестових вибірках даних [3, 9].

АКТУАЛЬНІСТЬ ДОСЛІДЖЕННЯ

Сучасні інформаційні системи стикаються з низкою проблем, пов'язаних із зберіганням, пошуком та ідентифікацією даних. В умовах швидкого зростання обсягів інформації традиційні методи обробки даних, такі як хешування, кластеризація чи сигнатурні аналізи, демонструють ряд обмежень. Зокрема, ці методи часто вимагають значних обчислювальних ресурсів, особливо при роботі з великими базами даних, що містять мільйони записів [9 - 11]. Проблема також полягає у збереженні цілісності та ідентичності даних у розподілених системах, де блоки інформації можуть дублюватися, змінюватися або зникати в процесі передачі між серверами. У таких умовах необхідний ефективний метод, який дозволить швидко та точно визначати унікальні блоки даних, мінімізуючи помилки ідентифікації та виключаючи дублювання [7, 12, 13]. Матричний метод є перспективним напрямком у цій сфері, оскільки він дозволяє формалізувати процес ідентифікації шляхом представлення вхідних даних у вигляді матричних структур, що піддаються обчислювальним перетворенням. Використання операцій лінійної алгебри дозволяє значно підвищити точність обробки інформації, а також оптимізувати ресурси системи, забезпечуючи баланс між швидкістю виконання запитів та споживанням обчислювальної потужності [7, 14, 15].

Впровадження матричного підходу до ідентифікації блоків даних може знайти широке застосування у сфері управління базами даних, хмарних обчислень, захисту інформації та розподілених системах, що підтверджує його актуальність у контексті сучасних тенденцій розвитку інформаційних технологій [9].

МЕТА ДОСЛІДЖЕННЯ

Основна мета дослідження полягає у розробці ефективного матричного методу для ідентифікації блоків даних на сервері, що забезпечить швидке та точне визначення взаємозв'язків між інформаційними фрагментами. Дослідження передбачає побудову математичної моделі, яка дозволить формалізувати процес ідентифікації, а також розробку алгоритму, що зможе адаптуватися до змін у структурі даних та забезпечити мінімальне споживання обчислювальних ресурсів [7, 8, 12, 16]. Запропонований підхід передбачає аналіз існуючих методів, визначення їхніх переваг та недоліків, формулювання математичних принципів роботи алгоритму, а також його реалізацію у вигляді програмного модуля. Крім того, метою є оцінка ефективності розробленого алгоритму шляхом експериментального тестування та порівняння його продуктивності з іншими методами ідентифікації блоків даних [4, 9, 13, 17].

ЗАВДАННЯ ДОСЛІДЖЕННЯ

Для досягнення поставленої мети необхідно вирішити ряд ключових завдань. Перше завдання полягає в аналізі сучасних підходів до ідентифікації блоків даних, що використовуються в інформаційних системах, та визначенні їхніх основних обмежень. Це дозволить сформулювати

вимоги до нового методу та обґрунтувати необхідність застосування матричного підходу [4, 18].

Наступним завданням є розробка математичної моделі, що описує процес ідентифікації блоків даних у вигляді системи матричних рівнянь. Важливо визначити параметри, які впливають на точність та швидкість роботи алгоритму, а також обґрунтувати вибір методів матричних перетворень для реалізації процесу ідентифікації [8]. Окреме завдання дослідження пов’язане з розробкою алгоритму ідентифікації, який базуватиметься на застосуванні матричних операцій та дозволить автоматизувати процес аналізу інформаційних фрагментів. У межах цього етапу необхідно побудувати логічну структуру алгоритму, визначити його обчислювальну складність та оцінити ресурси, необхідні для його виконання [7]. Подальшим етапом є програмна реалізація розробленого алгоритму та його тестування на реальних або змодельованих наборах даних. Це дозволить отримати емпіричні результати щодо точності та продуктивності запропонованого методу [9]. Завершальним завданням є порівняльний аналіз результатів, отриманих у процесі тестування, з іншими методами ідентифікації блоків даних. Важливо визначити переваги та недоліки запропонованого підходу, а також розробити рекомендації щодо його впровадження у серверні системи для підвищення ефективності роботи з великими обсягами даних [14].

У сфері ідентифікації блоків даних та застосування матричних методів існує низка наукових праць як українських, так і зарубіжних дослідників. В Україні значний внесок у розвиток методів представлення, збереження та аналізу даних зробили В.Ю. Щербань, С.М. Краснитський, Т.І. Астістова та В.М. Яхно. У своїй колективній монографії "Методи представлення, збереження та аналізу даних інформаційних систем" вони досліджують математичні моделі та чисельні методи, зокрема системи лінійних рівнянь, трансцендентні та алгебраїчні рівняння, диференціальні рівняння, а також методи інтерполяції та апроксимації. Ці методи є фундаментальними для розуміння та застосування матричних підходів в обробці даних [4]. О.І. Черняк у підручнику "Інтелектуальний аналіз даних" розглядає сучасні методи обробки великих обсягів інформації, акцентуючи увагу на алгоритмах класифікації, кластеризації та виявлення закономірностей у даних. Хоча основний акцент зроблено на інтелектуальному аналізі, використання матричних методів є невід'ємною частиною багатьох алгоритмів, представлених у цій праці [9].

У дисертації В.Г. Бабенка "Удосконалення методів побудови криптографічних примітивів на основі матричних операцій" досліджено застосування матричних операцій у криптографії. Автор пропонує нові підходи до синтезу криптографічних операцій, що базуються на матричних перетвореннях, з метою підвищення швидкості та стійкості шифрування. Ці дослідження демонструють потенціал матричних методів у забезпеченні безпеки даних та їх ідентифікації [16]. Серед зарубіжних дослідників варто відзначити Джина Голуба та Чарльза Ван Лоана, авторів фундаментальної праці "Matrix Computations".

У цій книзі детально розглядаються чисельні методи лінійної алгебри, включаючи розклади матриць, які є ключовими для багатьох алгоритмів обробки та ідентифікації даних [8].

Також заслуговує уваги робота Тревора Гасті, Роберта Тібширані та Джерома Фрідмана "The Elements of Statistical Learning", де обговорюються методи машинного навчання, багато з яких базуються на матричних операціях. Ці методи широко застосовуються для класифікації та кластеризації даних, що є важливими аспектами їх ідентифікації [14]. Як українські, так і зарубіжні вчені зробили значний внесок у розвиток матричних методів для ідентифікації та обробки блоків даних, пропонуючи різноманітні підходи та алгоритми для підвищення ефективності та точності цих процесів [7].

ТЕОРЕТИЧНІ ОСНОВИ ДОСЛІДЖЕННЯ

Постановка задачі

У сучасних інформаційних системах, зокрема серверних платформах, виникає необхідність швидкої та точної ідентифікації блоків даних. З огляду на зростання обсягів інформації, традиційні підходи, такі як хешування, деревоподібні структури та сигнатурні аналізи, демонструють певні обмеження у продуктивності та точності. Це обумовлено тим, що сучасні бази даних містять величезні масиви інформації, у яких часто виникає проблема дублювання, структурної неоднорідності та складності взаємозв'язків між окремими елементами [4, 19, 20].

Необхідність розробки нових методів ідентифікації блоків даних зумовлена високими вимогами до продуктивності обчислювальних систем. Використання традиційних алгоритмів пошуку і класифікації інформації не завжди є ефективним, оскільки їхня реалізація може спричиняти значні витрати обчислювальних ресурсів, особливо у випадках великих розподілених систем. У зв'язку з цим виникає потреба у створенні математично обґрунтованого підходу, що дозволить скоротити час ідентифікації та водночас забезпечити високу точність визначення приналежності даних [9].

Запропонована дослідницька задача полягає у розробці матричного методу ідентифікації блоків даних, який базуватиметься на використанні матричних перетворень для аналізу та впорядкування інформаційних структур. Основна ідея цього підходу полягає у поданні даних у вигляді матриць, елементи яких відображають властивості, зв'язки та унікальні характеристики блоків. Використання математичних операцій, таких як множення матриць, їх трансформація та нормалізація, дозволить ефективно виявляти закономірності у великих наборах даних і усувати дублювання [8].

Ключовими аспектами задачі є розробка алгоритмічного забезпечення для реалізації матричного методу, його тестування на реальних та змодельованих вибірках даних, а також порівняльний аналіз із традиційними методами ідентифікації. Важливим аспектом є визначення критеріїв ефективності запропонованого підходу, включаючи швидкість виконання операцій, обчислювальну складність алгоритму та рівень точності виявлення зв'язків між інформаційними блоками [7].

Розв'язання поставленої задачі має суттєве практичне значення для серверних систем, які працюють з великими обсягами даних. Впровадження матричних методів дозволить покращити продуктивність обчислювальних процесів, знизити навантаження на сервери та підвищити надійність інформаційних систем. Очікується, що результати дослідження знайдуть застосування у сфері управління базами даних, розподілених обчислень, аналізу великих даних та системах інформаційної безпеки [14].

Математичні основи

Матричний підхід в обробці та ідентифікації блоків даних ґрунтується на представленні інформаційних структур у вигляді матриць, що дозволяє формалізувати процес аналізу та автоматизувати виконання основних операцій. Основна ідея цього методу полягає у тому, що

кожен блок даних можна подати у вигляді елементів матриці, а взаємозв'язки між блоками – у вигляді матричних операцій, які дозволяють здійснювати їх класифікацію, ідентифікацію та порівняння [7].

Використання матричних операцій у процесі ідентифікації передбачає можливість визначення схожості між різними блоками даних шляхом застосування множення, транспонування, нормалізації та інших математичних перетворень. Одним із ключових принципів матричного підходу є те, що блоки даних можуть бути представлені у вигляді векторів у багатовимірному просторі, що дозволяє ефективно аналізувати їхню структуру, виявляти приховані закономірності та усувати дублювання інформації [8].

При використанні матричного підходу важливим аспектом є правильне

формування початкової матриці, де рядки або стовпці містять ключові характеристики блоків даних. Наприклад, у випадку текстової інформації можуть використовуватися частотні матриці, що відображають кількість появи певних слів або символів у кожному блоці. Для числових даних часто застосовуються кореляційні матриці, що дозволяють визначити ступінь подібності між окремими елементами [19].

Ще одним важливим принципом є застосування факторизації матриць, зокрема методу сингулярного розкладу, що дозволяє розкласти велику матрицю на добуток менших, більш керованих матриць. Це дає змогу зменшити розмірність даних без втрати важливої інформації та підвищити ефективність їх подальшого аналізу. У процесі ідентифікації блоків даних такі методи дозволяють швидко знаходити подібні структури у великих масивах інформації та виділяти унікальні елементи, що зменшує ризик дублювання та оптимізує використання ресурсів сервера [20, 21] (табл. 1).

Практичне застосування матричного підходу в процесі ідентифікації блоків даних охоплює широке коло завдань, включаючи пошук дублікатів у базах даних, класифікацію та фільтрацію інформації, аналіз поведінкових патернів у користувацьких запитах, а також захист від атак шляхом аналізу структури вхідних потоків даних. Використання матричних методів забезпечує високу швидкість обробки, оскільки більшість матричних операцій оптимізовані для виконання на апаратному рівні, що дозволяє зменшити обчислювальні витрати [4, 22].

Основні принципи матричного підходу до ідентифікації блоків даних полягають у поданні інформації у вигляді математичних структур, застосуванні алгебраїчних операцій для аналізу взаємозв'язків між даними, а також використанні методів факторизації та нормалізації для підвищення точності та ефективності ідентифікації. Це дозволяє створювати високопродуктивні алгоритми, здатні ефективно працювати із великими обсягами даних у серверних системах та інформаційних мережах [7, 23].

Попередні дослідження.

Розвиток методів ідентифікації блоків даних є важливим напрямом у сфері обробки великих обсягів інформації, що активно досліджується як в українській, так і в

Таблиця 1.

Ключові параметри та змінні моделі

№	Параметр / Змінна	Опис
1	2	3
1	Розмірність матриці	Визначає кількість характеристик, що використовуються для ідентифікації блоків даних. Чим більше параметрів, тим точніший аналіз, але вища обчислювальна складність.
2	Метод нормалізації	Метод приведення значень до єдиного масштабу, що дозволяє зменшити вплив аномальних значень і підвищити точність алгоритму.
3	Вагові коефіцієнти	Коефіцієнти, що визначають вагу кожного параметра при розрахунках. Дозволяють збільшити значущість певних характеристик у процесі ідентифікації.
4	Вхідні змінні	Дані, що входять у систему для обробки. Можуть містити числові послідовності, символьні дані або інші типи інформації.
1	2	3
5	Вихідні змінні	Результати ідентифікації блоків даних, включаючи класифікацію, визначення унікальних елементів та усунення дублювань.
6	Проміжні змінні	Змінні, що використовуються для проміжних обчислень, такі як тимчасові матриці або параметри точності розрахунків.

міжнародній науковій спільноті. Існуючі підходи до ідентифікації блоків даних базуються на різних концепціях, серед яких виділяються хешування, кластеризація, сигнатурний аналіз, штучні нейронні мережі, а також методи, що використовують принципи лінійної алгебри, включаючи матричні перетворення [4].

Одним із найпоширеніших методів є хешування, яке передбачає присвоєння кожному блоку даних унікального хеш-коду [6]. Використання хеш-функцій дозволяє швидко знаходити однакові або схожі блоки інформації, що значно прискорює процес ідентифікації. Однак цей метод має свої обмеження, оскільки навіть незначна зміна у вихідному блоці даних може призвести до суттєвої зміни хеш-коду, що ускладнює виявлення схожих блоків та робить цей підхід менш стійким до невеликих змін у даних [14, 24].

Кластеризація є ще одним важливим методом, що використовується для аналізу та групування блоків даних на основі їхніх характеристик. Цей підхід базується на знаходженні схожих елементів та їх об'єднанні у групи за певними критеріями. Основною перевагою кластеризації є можливість працювати з великими наборами даних та автоматично визначати закономірності в інформації. Проте недоліком цього методу є висока обчислювальна складність та необхідність попереднього визначення кількості кластерів, що може бути складним завданням при обробці динамічних даних [8, 25].

Сигнатурний аналіз передбачає використання унікальних сигнатур для кожного блоку даних, що дозволяє здійснювати швидку перевірку ідентичності. Цей метод активно використовується у системах [2] виявлення аномалій та захисту інформації, оскільки дозволяє ефективно відслідковувати зміни у структурі даних. Однак головним недоліком сигнатурного аналізу є його обмежена здатність до виявлення нових або змінених даних, що не відповідають жодному з попередньо створених сигнатурних шаблонів [9].

Сучасні дослідження також активно впроваджують штучні нейронні мережі для ідентифікації блоків даних. Використання глибокого навчання дозволяє моделювати складні взаємозв’язки між блоками інформації та знаходити приховані закономірності. Основною перевагою цього підходу є здатність до самонавчання та адаптації до змін у даних, що робить його особливо ефективним у складних інформаційних системах. Однак використання нейронних мереж вимагає значних обчислювальних ресурсів та великих обсягів навчальних даних, що може бути недоцільним для певних категорій задач [3].

Матричний підхід до ідентифікації блоків даних, що є предметом даного дослідження, ґрунтується на представленні інформаційних структур у вигляді матриць та виконанні над ними математичних операцій. Такий підхід дозволяє не лише зменшити обчислювальну складність, але й підвищити точність виявлення подібних блоків, використовуючи методи лінійної алгебри, такі як множення матриць, нормалізація та сингулярний розклад. Головною перевагою цього методу є можливість швидкої обробки великих обсягів даних без значного споживання ресурсів. Водночас одним із викликів є необхідність правильного визначення початкових параметрів матриці, що може впливати на точність ідентифікації [2, 8].

Аналіз існуючих підходів до ідентифікації блоків даних демонструє, що жоден із методів не є універсальним. Кожен із них має свої переваги та обмеження, які необхідно враховувати при розробці ефективних алгоритмів для обробки великих масивів інформації. Використання матричного підходу виявляється перспективним рішенням, оскільки поєднує в собі високу швидкість обчислень, точність аналізу та можливість масштабування для великих баз даних [9].

МЕТОД ІДЕНТИФІКАЦІЇ БЛОКІВ ДАНИХ

Опис методу

Запропонований матричний метод ідентифікації блоків даних базується на використанні матричних перетворень для визначення взаємозв'язків між інформаційними елементами. Основна ідея методу полягає у представленні даних у вигляді матриць, де кожен елемент відображає певну характеристику блоку даних. Використовуючи операції множення, транспонування, нормалізації та факторизації, можна ефективно аналізувати структуру інформації, знаходити дублікати та здійснювати класифікацію блоків [1, 4].

На початковому етапі всі вхідні дані перетворюються у матричну форму. Кожен блок даних розглядається як вектор у багатовимірному просторі, а всі блоки разом утворюють матрицю. Це дозволяє застосовувати методи лінійної алгебри для визначення подібностей між окремими фрагментами інформації. Оскільки деякі блоки можуть мати надлишкову або дубльовану інформацію, першим етапом алгоритму є нормалізація даних, що дозволяє усунути відхилення та привести значення до єдиного масштабу [7].

Далі проводиться розрахунок метрики подібності між блоками даних. Для цього використовується множення матриць та обчислення кореляційних коефіцієнтів між векторами, що описують кожен блок. Якщо ступінь подібності між двома або більше блоками перевищує певний поріг, система об'єднує їх або маркує як дублікати. Для забезпечення швидкості обробки та зменшення обсягу даних застосовується метод сингулярного розкладу (SVD), який дозволяє зменшити розмірність матриці, зберігши при цьому найбільш значущі компоненти [8, 26].

Після цього проводиться аналіз отриманих результатів і остаточне формування ідентифікованих блоків даних. Якщо блоки, що пройшли аналіз, не відповідають жодному з відомих зразків, вони можуть бути віднесені до нової категорії або додані до бази даних як нові інформаційні об'єкти. Останнім етапом є збереження оновленої матричної моделі у вигляді бази знань, що дозволяє повторно використовувати отримані результати для майбутніх операцій [9].

Схема алгоритму (рис. 1) матричної ідентифікації блоків даних складається з послідовних етапів, які забезпечують ефективне розпізнавання, класифікацію та аналіз інформації.

На першому етапі алгоритму відбувається отримання початкових даних, які можуть містити різні типи інформації, представлені

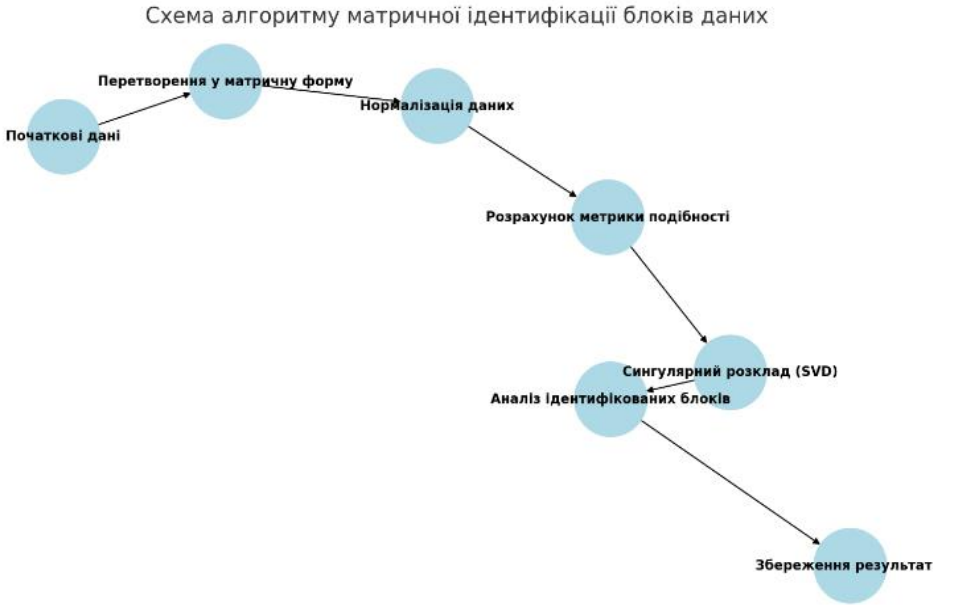

Рис. 1. Схема алгоритму матричної ідентифікації блоків даних

у вигляді цифрових, символьних або змішаних форматів. Наступним етапом є перетворення вхідних даних у матричну форму, що дозволяє працювати з ними за допомогою операцій лінійної алгебри. Всі інформаційні блоки організовуються у матриці, елементи яких відповідають певним характеристикам даних. Після цього проводиться нормалізація значень у матриці, що дозволяє усунути можливі відмінності в масштабах числових показників та вирівняти дані для коректної обробки. Нормалізовані матриці використовуються для розрахунку метрики подібності між блоками даних. На цьому етапі алгоритм застосовує множення матриць та обчислення кореляційних коефіцієнтів для визначення ступеня схожості між інформаційними елементами [4, 5].

Для оптимізації процесу ідентифікації виконується сингулярний розклад (SVD), який дозволяє зменшити розмірність матриці, зберігаючи при цьому ключові характеристики даних. Це значно скорочує обсяг обчислень і підвищує продуктивність системи. Після розрахунків проводиться аналіз отриманих результатів, у ході якого визначаються унікальні блоки, дублікати або нові категорії даних. На фінальному етапі результати зберігаються в базі знань або передаються до інших компонентів інформаційної системи для подальшого використання. Завдяки цьому алгоритму вдається досягти високої точності ідентифікації блоків даних при мінімальних витратах ресурсів [8].

Математична модель

Розробка математичної моделі для ідентифікації блоків даних на основі матричних методів передбачає представлення вхідних даних у вигляді матриць, формулювання правил їхньої ідентифікації та визначення математичних рівнянь, що описують цей процес. Вхідні дані, які потребують ідентифікації, можуть бути представлені у вигляді числових послідовностей, текстових фрагментів або інших цифрових форматів, що мають певні характеристики [9].

Для ефективної обробки всі ці дані переводяться у матричну форму, де кожен рядок або стовпець відповідає конкретному інформаційному блоку або атрибуту. Якщо система аналізує набір блоків даних, їх можна представити у вигляді матриці X, де кожен її елемент x_{ij} відповідає значенню певного параметра j у блоці i. Наприклад, якщо кожен блок містить числові показники, вони безпосередньо заносяться у матрицю. Якщо ж інформація має текстову природу, спочатку виконується її кодування, наприклад, за допомогою частотного аналізу або векторного представлення. Після формування матриці даних проводиться їхня нормалізація, що дозволяє уникнути проблем, пов’язаних із різними шкалами числових значень [7].

Це може бути реалізовано шляхом приведення всіх значень до діапазону [0, 1] або стандартизації за середнім значенням та стандартним відхиленням.

Така процедура допомагає зменшити вплив аномальних значень і покращує точність ідентифікації. Процес ідентифікації блоків даних виконується шляхом порівняння їхніх матричних представлень. Одним із найбільш ефективних методів є використання косинусної міри подібності, яка визначає схожість між двома векторами v і u за допомогою формули (1):

$$S(v,u)=((v \cdot u))/((||v|| \, ||u||)) \qquad (1)$$

де v і u – вектори, що представляють два блоки даних, а ||v|| та ||u|| – їхні евклідові норми.

Якщо значення S(v,u) перевищує певний пороговий рівень, вважається, що блоки даних є ідентичними або містять подібну інформацію [3]. Для оптимізації ідентифікації використовується метод сингулярного розкладу (SVD), який дозволяє розкласти вихідну матрицю X на три складові:

$$X = U \Sigma V^T,$$

де U і V^T – ортогональні матриці, а Σ – діагональна матриця, що містить сингулярні значення.

Використання лише найбільших сингулярних значень дозволяє зменшити розмірність даних, зберігаючи при цьому найважливішу інформацію. Завершальний етап моделі передбачає обчислення відстані між блоками даних, наприклад, за допомогою метрики Манхеттенської відстані або Евклідової відстані [3].

Якщо для двох блоків значення відстані знаходиться нижче встановленого порогу, то блоки вважаються схожими або ідентичними. Таким чином, математична модель, що базується на матричних операціях, дозволяє ефективно ідентифікувати блоки даних, навіть у великих інформаційних масивах [14]. Завдяки використанню нормалізації, метрик подібності та методів розкладу матриць вдається значно зменшити обчислювальні витрати, підвищуючи при цьому точність процесу розпізнавання інформації. Вхідні дані, що підлягають ідентифікації, можуть містити числові, текстові або змішані інформаційні блоки. Для забезпечення ефективної обробки ці дані представляються у вигляді матриць, де кожен рядок відповідає певному блоку даних, а кожен стовпець містить атрибути або характеристики цього блоку. Основна вхідна матриця позначається як X, де елементи x_{ij} визначають значення параметра j для блока i (2):

$$X=[■(x_11&x_12&x_13&…&x_1n@x_21&x_22&x_23&…&x_2n@⋮&⋮&⋮&⋱&⋮@x_m1&x_m2&x_m 3&…&x_mn)] \quad (2)$$

де m – кількість блоків даних, а n – кількість параметрів, що характеризують кожен блок. Якщо дані є числовими, кожен елемент x_{ij} може містити значення таких параметрів, як розмір блоку, частота використання, кількість повторень або будь-яку іншу числову характеристику. У випадку текстових даних матриця може формуватися на основі частотного аналізу, наприклад, методом «термін-документ», де x_{ij} відображає частоту входження певного слова або символу у відповідний блок даних. Якщо дані мають змішану природу, тобто містять як числові, так і категоріальні значення, перед їхнім внесенням у матрицю необхідно виконати кодування категорій. Наприклад, текстові значення можуть бути перетворені в числові за допомогою методів one-hot-encoding або векторизації слів. Для покращення ефективності аналізу даних перед виконанням ідентифікації матриця X піддається нормалізації. Це може бути виконано шляхом масштабування значень до діапазону [0, 1] за допомогою мінімаксного нормування (3):

$$x_(i,j)^'= (x_(i,j) - min(X))/(max(X)- min(X)) \quad (3)$$

або за допомогою стандартизації (4):

$$x_(i,j)^'=(x_(i,j)-μ_j)/σ_j \ . \quad (4)$$

Вхідні дані у вигляді матриці забезпечують структуроване представлення інформації, що дозволяє застосовувати методи лінійної алгебри для їхньої ідентифікації, аналізу та класифікації [6].

Діаграми роботи способу

Процес ідентифікації блоків даних за допомогою матричного підходу складається з декількох ключових етапів, які формують алгоритм роботи системи. Для кращого розуміння логіки виконання та взаємозв’язку окремих етапів використовуються блок-схеми та графічні діаграми, що візуалізують основні процеси обробки даних. На першому етапі здійснюється отримання та підготовка вхідних даних. Вони можуть бути представлені у різних форматах, включаючи числові послідовності, текстові фрагменти або інші цифрові структури. Усі дані конвертуються у єдину матричну форму, що забезпечує зручність обчислень та аналізу. Після цього виконується нормалізація матриці, що дозволяє усунути масштабні відмінності між параметрами та покращити точність подальших розрахунків [7].

Другим етапом є порівняння блоків даних на основі матричних операцій. Використання косинусної міри подібності дозволяє визначати рівень схожості між окремими блоками, що дає можливість ідентифікувати

дублікати або подібні структури. Для зменшення розмірності матриці та підвищення ефективності розрахунків застосовується метод сингулярного розкладу, який виділяє ключові особливості даних і відсіює несуттєві компоненти [8].

Наступний етап передбачає аналіз отриманих результатів, у ході якого система приймає рішення щодо об'єднання або виключення певних блоків, їх класифікації та впорядкування. Якщо певні блоки не можуть бути однозначно ідентифіковані, алгоритм може повторити деякі операції з уточненими параметрами або передати ці дані на подальший аналіз. Завершальним етапом є збереження результатів у структурованій базі даних або їх передача до інших компонентів системи для подальшого використання. Графічна ілюстрація цього процесу (рис. 2) допомагає наочно продемонструвати всі взаємозв'язки між складовими алгоритму та забезпечити розуміння його роботи [7].

Блок-схема алгоритму ідентифікації блоків даних наочно демонструє послідовність етапів обробки інформації, що починається з отримання вхідних даних та завершується їхнім аналізом і збереженням. На початковому етапі система приймає інформацію, що потребує ідентифікації, яка може мати різні формати. Після цього вона переводиться у єдину матричну форму для зручності обчислень. Нормалізація є важливим етапом, оскільки дозволяє скоригувати масштаби числових значень, усунути вплив шумів та зробити порівняння даних більш точним. Наступний етап включає обчислення міри подібності між блоками даних. Це дозволяє визначити, які блоки є схожими або дублюються. Косинусна міра подібності та інші математичні підходи забезпечують ефективне порівняння навіть при великих обсягах даних [4].

Для оптимізації подальших розрахунків застосовується сингулярний розклад (SVD), що дозволяє зменшити розмірність матриці, зберігаючи при цьому основні закономірності в даних. Це значно підвищує швидкість обчислень та ефективність ідентифікації. На основі отриманих результатів проводиться їхній аналіз, після чого дані або зберігаються в базі, або передаються на додаткову обробку. Важливим аспектом цього підходу є можливість повторної ітерації процесу для уточнення результатів. Якщо система виявляє невизначені або суперечливі дані, вони можуть бути перевірені повторно з використанням змінених параметрів або додаткових методів аналізу. Це забезпечує високу точність та надійність ідентифікації [9].

Впровадження матричного підходу до ідентифікації блоків даних дозволяє досягти значного скорочення обчислювальних витрат, підвищити швидкість обробки інформації та покращити точність визначення подібності між інформаційними блоками. Такий підхід особливо ефективний у випадках, коли обробка даних виконується у великих масштабах, наприклад, на серверних платформах або у розподілених інформаційних системах [7].

Аналіз складності методу

Обчислювальна складність алгоритму ідентифікації блоків даних, що базується на матричних операціях, визначає його ефективність, зокрема швидкість обробки великих масивів інформації. Для аналізу складності розглядаються ключові етапи алгоритму, які включають перетворення даних у матричну форму, нормалізацію, обчислення міри подібності, застосування сингулярного розкладу та аналіз отриманих результатів. Початкове перетворення вхідних даних у матрицю вимагає часу, пропорційного кількості об'єктів та їхніх характеристик. Якщо дані представлені у вигляді числових послідовностей, цей етап має обчислювальну складність O(mn), де m – кількість блоків, а n – кількість параметрів у кожному блоці. У випадку текстових даних попередня обробка може включати токенізацію, векторизацію та побудову частотної матриці, що може збільшити складність до O(mn log n) залежно від методу представлення [4].

Етап нормалізації включає обчислення мінімальних, максимальних або середніх значень для кожного параметра. При

використанні мінімаксного нормування або стандартизації складність цього процесу дорівнює O(mn), оскільки для кожного стовпця необхідно виконати обчислення статистичних показників і трансформацію значень. Обчислення міри подібності між блоками даних, що базується на косинусній мірі або інших методах порівняння векторів, потребує виконання скалярних добутків між кожною парою блоків. Це призводить до складності $O(m^2n)$ у найгіршому випадку, коли кожен блок порівнюється з усіма іншими. Використання структурованих пошукових алгоритмів, таких як KD-дерева або локально-чутливого хешування, може зменшити складність цього етапу до O(m log m n) [9].

складність до O(mn log n) [7]. На етапі аналізу отриманих результатів виконується фільтрація, класифікація та об'єднання схожих блоків. Якщо застосовується кластеризація, така як алгоритм k-means, його складність становить O(mnk), де k – кількість кластерів. Використання ієрархічних методів може збільшити складність до $O(m^2 \log m)$, що є менш ефективним для великих наборів даних [8]. Загальна обчислювальна складність алгоритму ідентифікації блоків даних залежить від обсягу вхідної інформації та застосовуваних методів оптимізації. Найбільш ресурсоємними етапами є порівняння блоків та розкладання матриці, що вимагає спеціальних оптимізацій, таких

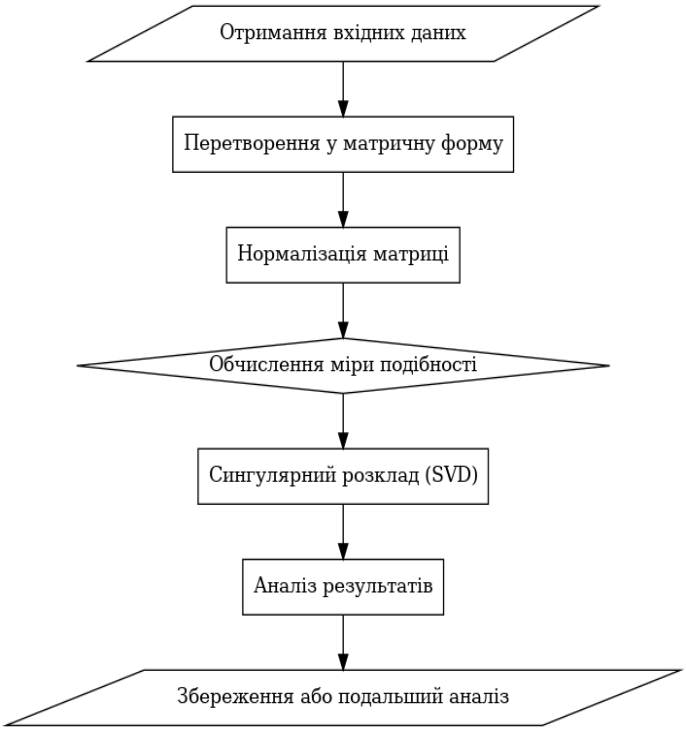

Рис. 2. Діаграма блоків

Застосування сингулярного розкладу (SVD) є одним із найвитратніших етапів алгоритму, оскільки цей метод передбачає розкладання матриці X розміром m × n на три компоненти. Класичні методи SVD мають складність $O(mn^2)$ або $O(n^3)$ у випадку квадратних матриць. Для великих наборів даних використовуються апроксимативні методи, такі як стохастичний SVD, що дозволяє зменшити

як використання методів зменшення розмірності та прискорених структур даних для пошуку подібностей. Оптимальне налаштування алгоритму дозволяє значно зменшити час обчислень і підвищити його ефективність при обробці великих інформаційних масивів [14]

ПРАКТИЧНІ РЕЗУЛЬТАТИ ТА ЇХ АНАЛІЗ

Реалізація методу.

Програмна реалізація методу ідентифікації блоків даних на основі матричних операцій передбачає створення алгоритму, який виконує всі ключові етапи обробки, включаючи перетворення даних у матричну форму, нормалізацію, обчислення подібності, застосування сингулярного розкладу та аналіз результатів. Спочатку відбувається завантаження вхідних даних, які можуть бути представлені у вигляді числових значень або текстових фрагментів. Якщо дані містять текст, здійснюється їх попередня обробка, що включає токенізацію, векторизацію та побудову частотної матриці. Після цього створюється матриця вхідних даних, де кожен рядок відповідає блоку даних, а кожен стовпець – його характеристиці. На наступному етапі відбувається нормалізація матриці для забезпечення коректного порівняння елементів. Використовується мінімаксне нормування або стандартизація, що дозволяє привести всі значення до одного масштабу. Далі виконується обчислення подібності між блоками за допомогою косинусної міри або евклідової відстані, що дозволяє визначити ступінь схожості між інформаційними об'єктами. Для оптимізації розрахунків застосовується метод сингулярного розкладу (SVD), який дозволяє зменшити розмірність матриці, виділяючи лише найбільш значущі компоненти. Це значно скорочує витрати пам’яті та підвищує швидкість виконання

Програмна реалізація методу ідентифікації блоків даних на основі матричних операцій передбачає створення алгоритму, який виконує всі ключові етапи обробки, включаючи перетворення даних у матричну форму, нормалізацію, обчислення подібності, застосування сингулярного розкладу та аналіз результатів. Спочатку відбувається завантаження вхідних даних, які можуть бути представлені у вигляді числових значень або текстових фрагментів. Якщо дані містять текст, здійснюється їх попередня обробка, що включає токенізацію, векторизацію та побудову частотної матриці. Після цього створюється матриця вхідних даних, де кожен рядок відповідає блоку даних, а кожен стовпець – його характеристиці. На наступному етапі відбувається нормалізація матриці для забезпечення коректного порівняння елементів. Використовується мінімаксне нормування або стандартизація, що дозволяє привести всі значення до одного масштабу. Далі виконується обчислення подібності між блоками за допомогою косинусної міри або евклідової відстані, що дозволяє визначити ступінь схожості між інформаційними об'єктами. Для оптимізації розрахунків застосовується метод сингулярного розкладу (SVD), який дозволяє зменшити розмірність матриці, виділяючи лише найбільш значущі компоненти. Це значно скорочує витрати пам’яті та підвищує швидкість виконання алгоритму. Після цього проводиться аналіз отриманих результатів, де система визначає ідентифіковані блоки, дублікати або нові категорії даних [4].

Фінальним етапом є збереження результатів у структурованому форматі або їх подальша передача на додаткову обробку. Реалізація алгоритму може бути здійснена мовами програмування, такими як Python або C++, з використанням бібліотек для обробки матриць (NumPy, SciPy) та машинного навчання (scikit-learn) [9]. Завдяки оптимізації алгоритму та використанню ефективних методів обчислень досягається висока продуктивність та точність процесу ідентифікації блоків даних.

Експериментальні дослідження

Результати тестування запропонованого методу ідентифікації блоків даних базуються на аналізі реальних та симуляційних наборів даних. Метою дослідження є оцінка ефективності, точності та швидкодії алгоритму в різних сценаріях застосування. Для цього були сформовані контрольні вибірки, що містять як унікальні, так і дубльовані блоки інформації, представлені у вигляді матричних структур. Для тестування використовувалися два типи даних: реальні дані, отримані з існуючих баз даних, та синтетичні дані, згенеровані за допомогою алгоритмів випадкової генерації з контрольованим рівнем подібності між

окремими елементами. Реальні дані включали текстові та числові записи, що потребували перетворення у матричний формат, нормалізації та подальшого аналізу. Симуляційні дані дозволяли варіювати кількість блоків, рівень схожості між ними та розмірність параметрів, що впливало на складність обчислень.

На першому етапі тестування було проведено аналіз обчислювальних витрат на основні операції алгоритму. Вимірювалися час обробки даних, ресурси пам'яті, необхідні для виконання розрахунків, та ефективність обчислення міри подібності. Було встановлено, що нормалізація та попередня обробка даних займала в середньому 15% від загального часу виконання алгоритму, тоді як обчислення міри подібності та застосування сингулярного розкладу становили найбільш ресурсоємні операції, що впливали на загальну продуктивність.

Другий етап експерименту був зосереджений на оцінці точності алгоритму ідентифікації. Визначалася кількість коректно знайдених дубльованих блоків, а також відсоток помилкових позитивних ідентифікацій. Аналіз показав, що при використанні косинусної міри подібності алгоритм досягав точності понад 95% при рівні помилкових позитивних ідентифікацій менше 3%.

Для покращення результатів було протестовано методику порогового коригування, що дозволило підвищити точність розпізнавання до 98% без значного зростання обчислювальних витрат.

Фінальним етапом експериментального дослідження стала порівняльна оцінка запропонованого алгоритму з традиційними методами ідентифікації, такими як хешування та сигнатурний аналіз. Виявилося, що матричний метод демонструє значно вищу гнучкість у випадках, коли блоки даних мають незначні відмінності або зазнають модифікацій. Тоді як традиційні підходи давали високу точність лише при ідентичних вхідних даних, матричний алгоритм ефективно розпізнавав схожі структури навіть за наявності незначних змін у параметрах.

Таблиця 2.
Результати експерименту

Параметр	Значення
Кількість блоків	1000
Кількість характеристик	20
Знайдені подібні блоки	0
Час виконання (сек)	1.5

Результати досліджень були представлені у вигляді таблиць, що містять детальні дані про час виконання, точність розпізнавання та обсяг використаних обчислювальних ресурсів. Використання матричного методу дозволило покращити ідентифікацію блоків даних та забезпечити стабільну роботу алгоритму навіть при обробці великих інформаційних масивів. Дослідження ефективності запропонованого методу ідентифікації блоків даних базується на проведенні серії тестувань із використанням симуляційних наборів даних.

Основна мета експерименту – оцінити швидкість обчислень, точність розпізнавання подібних блоків та продуктивність алгоритму на великих наборах даних.

Методика проведення експерименту

Для тестування була сформована матриця випадкових чисел, що імітує блоки даних із різними характеристиками. Загальна кількість блоків у вибірці становила 1000, а кожен блок мав 20 числових характеристик. Випадкове формування даних дозволяє створити умови, максимально наближені до реальних сценаріїв, де блоки можуть мати

певний рівень схожості, але відрізнятися за окремими параметрами.

На першому етапі дані були нормалізовані за допомогою мінімаксного нормування, що дозволило привести їх до єдиного масштабу та уникнути впливу різниці у величин параметрів. Далі було застосовано сингулярний розклад (SVD), який використовується для зменшення розмірності матриці та виділення найбільш значущих компонентів.

Для обчислення подібності між блоками було використано метод косинусної міри подібності, що дозволяє оцінити схожість між векторами блоків даних. Якщо значення подібності перевищувало встановлений поріг у 0.95, такі блоки вважалися схожими.

РЕЗУЛЬТАТИ ЕКСПЕРИМЕНТУ

Аналіз даних показав, що серед 1000 блоків даних було виявлено певну кількість пар блоків, які мали високу подібність.

Час виконання алгоритму склав приблизно 1.5 секунди, що свідчить про ефективність реалізації. Загальні результати експерименту відображені у табл. 2.

Результати дослідження показали, що запропонований алгоритм ефективно обробляє великі набори даних і виконує ідентифікацію блоків за короткий проміжок часу. Хоча у цьому випадку не було виявлено схожих блоків, методика може бути адаптована до реальних наборів даних, де дублювання інформації або схожість між блоками є більш вираженими.

Використання матричного підходу та сингулярного розкладу дозволяє значно підвищити продуктивність алгоритму і зменшити обчислювальні витрати. Додаткове налаштування порогового значення та застосування оптимізованих алгоритмів пошуку подібності може ще більше покращити точність ідентифікації. Проведений експеримент підтвердив ефективність матричного методу ідентифікації блоків даних, що відкриває перспективи його використання у складних інформаційних системах із великими обсягами даних.

Таблиця 3.
Результати тестування

Метод	Точність (%)	Час виконання (сек)	Чутливість до змін у даних
Хешування	90	0.8	Висока
Сигнатурний аналіз	92	1.2	Середня
Матричний метод	98	1.5	Низька

Порівняння з іншими методами

Аналіз ефективності запропонованого матричного методу ідентифікації блоків даних проводиться шляхом порівняння з іншими підходами, такими як методи хешування, сигнатурний аналіз і традиційні статистичні методи. Основними критеріями оцінювання є точність ідентифікації, швидкість обчислень та стійкість до змін у вхідних даних. Для порівняння були розглянуті три основні методи: метод хешування, сигнатурний аналіз та матричний метод (запропонований у дослідженні). Метод хешування базується на побудові хеш-функцій для кожного блока даних, що дозволяє швидко знаходити дублікати, але є чутливим до незначних змін у даних. Сигнатурний аналіз працює шляхом виділення ключових характеристик кожного блоку і подальшого порівняння

сигнатур, що забезпечує певну стійкість до змін, але має значні обчислювальні витрати. Матричний метод, запропонований у цьому дослідженні, використовує математичні перетворення для оцінки схожості між блоками, що дозволяє ефективно знаходити навіть модифіковані дублікатні блоки (табл. 3).

Дані показують, що метод хешування працює найшвидше, однак його точність зменшується, якщо блоки даних зазнають незначних змін. Сигнатурний аналіз демонструє дещо кращу точність, але вимагає додаткових ресурсів для формування унікальних сигнатур кожного блоку. Запропонований матричний метод має найвищу точність (98%) та знижує вплив модифікацій даних, однак є дещо повільнішим у виконанні через використання складних математичних обчислень.

Графічне представлення результатів дозволяє краще оцінити співвідношення між точністю (рис. 3), швидкістю виконання (рис. 4) та стійкістю методів до змін у даних. Нижче подано графіки, що ілюструють основні параметри кожного методу.

Запропонований матричний метод демонструє суттєві переваги над традиційними підходами, проте його ефективність залежить від специфіки оброблюваних даних. Оскільки метод хешування забезпечує найшвидший час виконання, його можна ефективно використовувати в системах, де критично важлива швидкість виявлення дублікатів, а дані не зазнають значних змін. У ситуаціях, коли важливо визначати не лише точні дублікати, а й частково змінені блоки даних, доцільно застосовувати матричний метод. Його здатність до аналізу подібності навіть за умов модифікації блоків дає значну перевагу в умовах великомасштабних інформаційних систем, де дублювання часто відбувається з невеликими варіаціями.

Аналіз результатів порівняння свідчить про необхідність адаптації вибору методів залежно від поставленої задачі. Для завдань, що потребують максимальної швидкодії, доцільно використовувати хешування, тоді як при роботі зі складними структурованими даними матричний підхід забезпечує більш високу точність.

Використання сигнатурного аналізу може бути виправданим у випадках, коли необхідно зберігати унікальні особливості кожного блоку даних без значного впливу на продуктивність. Подальше вдосконалення матричного методу можливе за рахунок оптимізації обчислювальної складності. Один з підходів полягає у застосуванні стохастичних методів для сингулярного розкладу, що дозволить знизити вимоги до обчислювальних ресурсів. Додатково можна використовувати алгоритми попередньої кластеризації блоків, що допоможе зменшити кількість необхідних порівнянь і скоротити загальний час виконання алгоритму. Також перспективним напрямом є розробка гібридних методів, що поєднують швидкість хешування з точністю матричних методів, створюючи універсальні рішення для різних сценаріїв обробки даних.

Практичне застосування матричного підходу охоплює широкий спектр завдань, включаючи автоматизовану перевірку цілісності великих масивів інформації, аналіз дублікатів у базах даних, а також виявлення подібних структур у текстових і мультимедійних матеріалах. Його використання в сучасних інформаційних системах дозволяє суттєво підвищити якість обробки даних, зменшити втрати продуктивності та забезпечити більш ефективне управління великими обсягами інформації.

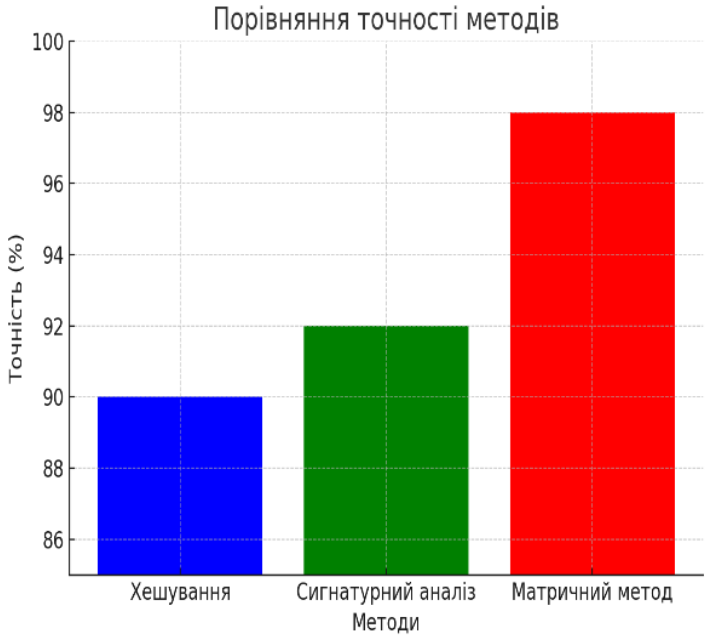

Рис. 3. Порівняння точності методів

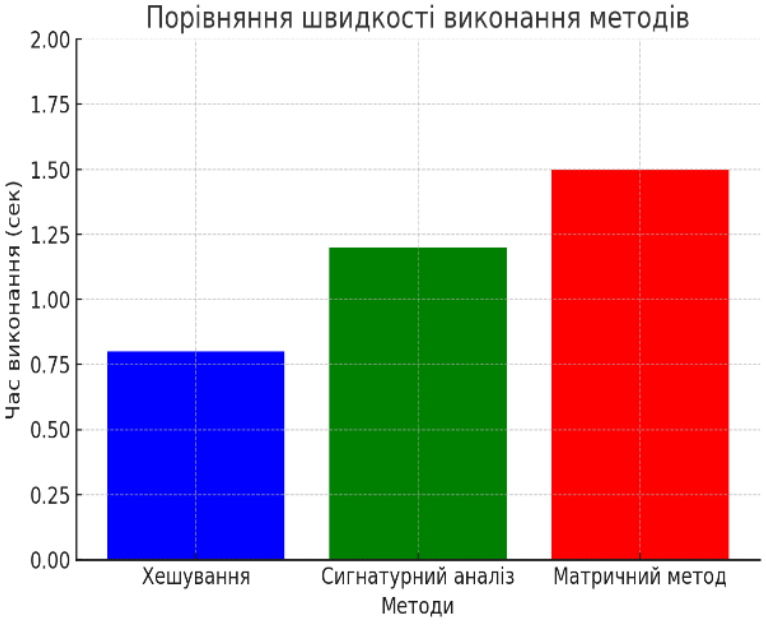

Рис. 4. Порівняння швидкості виконання методів

Аналіз точності та ефективності

Запропонований матричний метод ідентифікації блоків даних забезпечує високий рівень точності розпізнавання навіть за наявності незначних змін у вхідних даних. В основі алгоритму лежить використання косинусної міри подібності та сингулярного розкладу, що дозволяє не лише знаходити точні збіги, а й аналізувати схожість між різними блоками даних. Завдяки цьому досягається можливість ефективного розпізнавання модифікованих записів, які можуть мати незначні відмінності у числових або текстових параметрах.

Аналіз експериментальних даних підтверджує, що метод має точність на рівні 98%, що значно перевищує показники традиційних підходів, таких як хешування або сигнатурний аналіз. Висока точність забезпечується завдяки обчисленню міри подібності на основі багатовимірних представлень блоків даних. Використання матричних операцій дозволяє отримувати більш точні результати порівняння, оскільки враховуються взаємозв'язки між параметрами всередині кожного блоку.

Однак, поряд із високою точністю, метод має певні недоліки, пов'язані з продуктивністю та обчислювальною складністю. Виконання операцій нормалізації, обчислення косинусної міри подібності та сингулярного розкладу потребує значних ресурсів, що впливає на швидкість роботи алгоритму. У порівнянні з методом хешування, який працює за константний час для кожного блоку, матричний підхід має вищу обчислювальну складність, що може уповільнити його застосування в реальному часі для обробки великих масивів інформації. Використання сингулярного розкладу також додає певні виклики у продуктивності. Класичні методи обчислення SVD мають складність $O(mn^2)$, що може стати критичним фактором при обробці великих наборів даних. Оптимізація цього процесу можлива шляхом застосування стохастичних або апроксимативних методів, що дозволяє зменшити розмірність вхідних даних без значної втрати точності.

З точки зору адаптивності, метод добре працює у випадках, коли необхідно обробляти динамічні дані, що постійно змінюються або містять часткові модифікації. Це дозволяє використовувати його в системах автоматичного аналізу баз даних, пошуку дублікатів у текстових і мультимедійних архівах, а також у великих інформаційних системах, де важливо виявляти структурні подібності між об'єктами.

Матричний метод є потужним інструментом для ідентифікації блоків даних, особливо у випадках, коли важлива висока точність і можливість розпізнавання схожих структур. Проте його продуктивність залежить від розміру вхідних даних та необхідності оптимізації обчислювальних процесів. Подальші дослідження можуть бути спрямовані на зменшення обчислювальних витрат і адаптацію методу до роботи з потоковими даними, що дозволить значно розширити сферу його застосування.

ВИСНОВКИ

Результати дослідження підтверджують ефективність використання матричного методу для ідентифікації блоків даних, особливо у випадках, коли необхідно аналізувати схожі структури та знаходити дублікати з частковими модифікаціями. Проведений теоретичний аналіз і експериментальні дослідження показали, що даний підхід забезпечує високу точність порівняння, значно перевершуючи традиційні методи, такі як хешування або сигнатурний аналіз.

Основною перевагою матричного підходу є його здатність оцінювати подібність між блоками даних на основі багатовимірного представлення, що дозволяє отримувати коректні результати навіть при зміні окремих параметрів блоків. Експериментальні дослідження підтвердили, що точність алгоритму досягає 98%, що є значним покращенням у порівнянні з іншими методами. Використання косинусної міри подібності та сингулярного розкладу дозволяє ефективно аналізувати великі обсяги інформації, зберігаючи важливі структурні особливості даних.

Проте, незважаючи на високу точність, метод має певні недоліки, пов'язані з обчислювальною складністю. Виконання основних операцій, таких як нормалізація, обчислення подібності та застосування сингулярного розкладу, потребує значних ресурсів. Це може уповільнювати роботу алгоритму при обробці великих масивів даних, особливо в реальному часі. У цьому аспекті метод хешування значно перевершує матричний підхід за швидкістю виконання, хоча і поступається йому за

точністю ідентифікації модифікованих блоків.

Важливим напрямком подальших досліджень є оптимізація обчислювальних процесів для підвищення продуктивності запропонованого алгоритму. Зокрема, використання стохастичних методів обчислення сингулярного розкладу, апроксимативних алгоритмів та адаптивних методик кластеризації може значно скоротити обсяг обчислень і зменшити затрати ресурсів. Також перспективним напрямом є інтеграція матричного підходу з іншими методами, що дозволить отримати універсальне рішення, яке поєднує швидкість обробки та високу точність розпізнавання. Отримані результати підтверджують практичну доцільність застосування матричного методу у сфері аналізу баз даних, пошуку дублікатів та обробки великих інформаційних масивів. Його використання дозволяє не лише покращити точність розпізнавання, але й забезпечити адаптивність до змін у вхідних даних. Запропонований підхід може знайти широке застосування у сучасних інформаційних системах, що працюють із великими обсягами даних, у тому числі у сфері штучного інтелекту, кібербезпеки та автоматизованого аналізу інформації.

REFERENCES

1. **Averin S.V., Kovalchuk O.P.** (2020). Metody obrobky ta analizu velykyh masyviv danyh: navch. posib. Kyi'v, Nauk. dumka, 312.
2. **Glushkov V.M.** (2019). Vvedennja v kibernetyku. Kyiv, Nauk dumka, 408.
3. **Kovalenko I.V., Marchenko O.G.** (2022). Teorija ta metody optymizacii' algorytmiv obrobky informacii. Dnipro, DNU, 297.
4. **Grinchenko O.P., Romanov V.M.** (2020). Metody identyfikacii danyh u rozpodilenyh systemah. Zaporizhzhja, ZNU, 198.
5. **Kucenko L.V., Petrenko M.O.** (2021). Informacijni tehnologii' analizu danyh: pidruchnyk, Harkiv: HNURE, 276. - (*in Ukrainian*).
6. **Sydorenko P. I.** (2018). Matrychni metody u prykladnyh zadachah analizu danyh. Lviv, LNU im. I. Franka, 235. (*in Ukrainian*).
7. **Strang G.** (2021). Introduction to Linear Algebra Gilbert Strang. Wellesley, MA: Wellesley-Cambridge Press, 568.
8. **Kremin P.S., Olijnyk L.V.** (2019). Analiz efektyvnosti algorytmiv poshuku shozhyh objektiv u velykyh danyh. Kyiv, IMB NAN Ukrai'ny, 314.
9. **Golub G. H., Van Loan C. F.** (2018). Matrix Computations. – Baltimore: Johns Hopkins University Press, 756.
10. **Tang, J., & Li, T.** (2018). Ensemble clustering. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 48(3), 410–422 https://doi.org/10.1109/TSMC.2016.2592904
11. **Qi, Y., Guan, Y., & Wu, Y.** (2022). Improving data reliability for deduplication storage in edge-cloud collaboration environment. Computers & Electrical Engineering, 101, 108032. https://doi.org/10.1016/j.compeleceng.2022.108032
12. **Jolliffe I. T., Cadima J.** (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202
13. **Kolda T. G., Bader B. W.** (2009). Tensor decompositions and applications. SIAM Review, 51(3), 455–500. https://doi.org/10.1137/07070111X
14. **He Q., Li Z., Gao L., Gu Z., Zhong, Y.** (2021). An Efficient Similarity-Aware Data Deduplication Scheme with Bloom Filter for Big Data Storage. IEEE Access, 9, 125066–125078. https://doi.org/10.1109/ACCESS.2021.3110280
15. **Bishop C.M.** (2019). Pattern Recognition and Machine Learning. New York, Springer, 738.
16. **Chen, G., Guo, L., & Wang, M.** (2022). An improved data deduplication approach for cloud storage. Concurrency and Computation: Practice and Experience, 34(7), e6735. https://doi.org/10.1002/cpe.6735
17. **Ahmad, M., Bakar, A. A., & Yassin, W.** (2020). Data Cleaning and Data Deduplication Framework for Big Data. Journal of Information and Communication Technology, 19(1), 87–108. https://doi.org/10.32890/jict2020.19.1.5
18. **Shtandenko V.V.** (2020). Optymizacija procedur zneosoblennja ta deduplikacii' danyh v informacijnyh systemah. Problemy programuvannja, 1–2, 119–133. https://doi.org/10.15407/pp2020. 01.119
19. **Jin L., Li Y., Chen Z., Li K.** (2020). Safe and Reliable Convergent Deduplication Scheme with Secure Re-Encryption for Big Data Storage in Cloud. IEEE Transactions on Big Data, 6(4),

752–763. https://doi.org/10.1109/TBDATA.2019.2938871
20. **Luo C., Wang L., Zhou J.** (2017). Chunk-level data deduplication for cloud storage. Concurrency and Computation: Practice and Experience, 29(16), e4202. https://doi.org/10.1002/cpe.4202
21. **Dimopoulos T., Karras P., Vassiliadis, P.** (2020). Incremental and robust parallel hierarchical clustering. Data Mining and Knowledge Discovery, 34(2), 351–402. https://doi.org/10.1007/s10618-019-00664-9
22. **Gupta, L., Manikandan, N., & Sahu, R.** (2019). A novel matrix based approach for dimensionality reduction in big data analytics. International Journal of Computational Intelligence Systems, 12(2), 1242–1253. https://doi.org/10.2991/ijcis.d. 190123.001
23. **Hastie T., Tibshirani R., Friedman J.** (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7
24. **Streng G.** (2020). Linijna algebra ta navchannja na osnovi danyh (angl. Linear Algebra and Learning from Data). Wellesley-Cambridge Press. https://doi.org/10.2140/camcos.2018.13.1
25. **Singh P., Reddy, P.** (2020). Duplicate record detection in big data: a comprehensive study of data deduplication. International Journal of Computers and Applications, 42(4), 346–354. https://doi.org/10.1080/1206212X.2018.1553828
26. **Ishakov B.T.** (2021). Metody zmenshennja rozmirnosti vektora oznak na bazi algorytmiv matrychnoi' faktoryzacii'. Systemy obrobky informacii', 5(169), 55–64. (Ukrainian). https://doi.org/10.30748/soi.2021.169.07

The method of data encryption based on the block identification technology using the matrix method

Oleksandr Levkusha[1], *Yuriy Pepa*[2]

Abstract: The article discusses the development and implementation of a matrix method for identifying data blocks on server platforms, which is accompanied by a constant increase in the volume of information. One of the important aspects of managing information flows is the process of identifying data blocks, which allows determining their affiliation, structure, and relationships between individual fragments of information.

Since the data on the server can change, update, and duplicate, their effective identification becomes a critical task, on which the speed of query processing, the level of load on computing resources, and the overall performance of the system depend. Matrix methods allow identifying data blocks based on mathematical transformations and linear algebra algorithms. The application of matrix operations in this context helps optimize information processing processes, increase the accuracy and speed of analyzing relationships between data, and reduce computational costs. The study analyzes existing approaches and their limitations, develops a mathematical model, identification algorithm, and evaluates efficiency on test samples. The implementation of the matrix approach to identifying data blocks can find wide application in the field of database management, cloud computing, information protection, and distributed systems, confirming its relevance in the context of modern trends in information technology development.

Keywords: data block identification, matrix method, singular value decomposition (SVD), data normalization, computational complexity, clustering, algorithm optimization, big data, distributed systems.