УДК 004.275:658.8
DOI https://doi.org/10.32689/maup.it.2024.4.18

Денис РЕДЬКО
аспірант кафедри інженерії програмного забезпечення та кібербезпеки, Київський національний торговельно-економічний університет, d.redko@knute.edu.ua
ORCID: *0009-0003-5827-264X*

Альона ДЕСЯТКО
доктор філософії «Комп'ютерні науки», доцент кафедри інженерії програмного забезпечення та кібербезпеки, Київський національний торговельно-економічний університет, desyatko@knute.edu.ua
ORCID: *0000-0002-2284-3418*

Байтума БІСАРІНОВ
доктор філософії «Комп'ютерні науки», кафедра інформаційних систем, Казахський національний університет імені Аль-Фарабі, Алматинський університет енергетики та телекомунікацій, baituma_bai@gmail.com
ORCID: *0000-0002-2218-0749*

Айгуль БІСАРІАНОВА
доктор філософії «Комп'ютерні науки», доцент кафедри інформаційних систем, Міжнародний університет інформаційних технологій (IITU), King's College London, aigulbis@gmail.com
ORCID: *0000-0001-6629-3051*

ОГЛЯД МЕТОДІВ АНАЛІЗУ ТРАФІКУ КОМПАНІЇ НА ОСНОВІ АНСАМБЛЕВОЇ КЛАСТЕРИЗАЦІЇ

Анотація*. У статті наведено огляд існуючих досліджень, присвячених аналізу трафіку компаній з використанням методів кластеризації даних. Виходячи з аналізу наукових публікацій, розглянутих у цій роботі, підкреслюється важливість розробки нових кібернетичних систем для аналізу трафіку у великих організаціях. Подібні системи, насамперед, мають бути спрямовані на оптимізацію маршрутизації, зниження витрат та підвищення швидкості доставки. Розглядається можливість створення подібної кластерної платформи, що забезпечує розподілену обробку даних та інтеграцію різних типів сховищ даних. Для ефективного збору та зберігання інформації пропонується використовувати такі джерела, як серверні логи, дані з датчиків трафіку, геолокаційні відомості та маршрути пересування користувачів, також залежно від специфіки бізнес-процесів компанії можуть використовуватися й інші дані.*

Метою дослідження *є систематизація існуючих методів та підходів до аналізу трафіку компаній з використанням різних методів кластеризації даних, і в тому числі колективних (ансамблевих) рішень. У статті застосовується* ***методологія*** *аналітичного методу дослідження, який включає огляд існуючої літератури, аналіз попередніх робіт і систематизацію знань в області кластерного аналізу трафіку. Наше дослідження фокусується на оцінці різних підходів та технологій, які використовуються для обробки великих даних.*

Наукова новизна *полягає у розгляді ансамблевих методів кластеризації для аналізу мережевого трафіку, що забезпечують масштабованість, швидкість обробки та гнучкість систем. Запропоновано інтеграцію різних джерел даних для оптимізації бізнес-процесів і прийняття рішень, враховуючи сучасні виклики Big Data.*

Висновки. *Було продемонстровано, що кластерні рішення, у тому числі, на основі колективних (ансамблевих) алгоритмів забезпечують масштабованість, високу швидкість обробки, доступність даних та сервісів, а також гнучкість у застосуванні різноманітних інструментів та технологій. Тим не менш, реалізація таких систем пов'язана з технічними викликами, що вимагають глибоких знань у галузі Big Data, машинного навчання та кластерних технологій, а також значних витрат на обладнання, програмне забезпечення та кваліфікованих фахівців. Незважаючи на ці складності, застосування колективних кластерних рішень для аналізу великих даних може забезпечити компаніям значні конкурентні переваги через оптимізацію бізнес-процесів, покращення якості прийняття рішень та підвищення загальної ефективності діяльності.*

Ключові слова: *мережевий трафік, великі дані, методи аналізу даних, кластеризація, колективні рішення, ансамблеві моделі.*

Denys REDKO, Alona DESIATKO, Baituma BISSARINOV, Aigul BISSARINOVA. OVERVIEW OF COMPANY TRAFFIC ANALYSIS METHODS BASED ON ENSEMBLE CLUSTERING

Annotation*. The article provides an overview of existing research devoted to the analysis of company traffic using data clustering methods. Based on the analysis of scientific publications reviewed in this work, the importance of developing new cybernetic systems for traffic analysis in large organizations is emphasized. Such systems, first of all, should be aimed at optimizing routing, reducing costs and increasing delivery speed. The possibility of creating a similar cluster platform that provides distributed data processing and integration of various types of data storage is under consideration. For effective collection and storage of information, it is suggested to use such sources as server logs, data from traffic sensors, geolocation information and user movement routes, and depending on the specifics of the company's business processes, other data may be used.*

The purpose of the study *is to systematize existing methods and approaches to the analysis of company traffic using various methods of data clustering, including collective (ensemble) solutions. The article uses* ***the methodology*** *of the analytical method of research, which includes a review of existing literature, analysis of previous works and systematization of knowledge in the field of traffic cluster analysis. Our research focuses on evaluating different approaches and technologies used to process big data.*

The scientific novelty *consists in the consideration of ensemble clustering methods for network traffic analysis, which provide scalability, processing speed and flexibility of systems. The integration of various data sources is proposed to optimize business processes and decision-making, taking into account the modern challenges of Big Data.*

Conclusions*. It was demonstrated that cluster solutions, including those based on collective (ensemble) algorithms, provide scalability, high processing speed, availability of data and services, as well as flexibility in the application of various tools and technologies. However, the implementation of such systems is associated with technical challenges that require deep knowledge in the field of Big Data, machine learning and cluster technologies, as well as significant costs for hardware, software and skilled professionals. Despite these difficulties, the application of collective cluster solutions for big data analysis can provide companies with significant competitive advantages by optimizing business processes, improving the quality of decision-making and increasing the overall efficiency of operations.*

Key words: *network traffic, big data, data analysis methods, clustering, collective solutions, ensemble models.*

Вступ. Використання технологій Big Data при вирішенні завдань, пов'язаних з аналізом мережевого трафіку для великих компаній, відіграє ключову роль, зокрема, при його оптимізації та масштабуванні структури мережі компанії, оскільки подібні технології дозволяють отримувати цінну інформацію з величезних масивів корпоративного трафіку, виявляти приховані закономірності та тенденції, що є критично важливим для покращення продуктивності та безпеки, у тому числі для мереж компаній. Ці дані необхідно структурувати, класифікувати та піддавати глибокому аналізу для того, щоб вирішити вищезазначені завдання [1].

Кластерний аналіз (або далі будемо використовувати абревіатуру КА) є основою для багатьох підходів до дослідження трафіку [3, 4]. Кластеризація, є процесом сегментації даних шляхом об'єднання схожих елементів у групи або «кластери», допомагає виявляти однорідні сегменти трафіку, які можуть бути проаналізовані як окремі одиниці з певними характеристиками [2]. В результаті КА формуються групи об'єктів з високим ступенем подібності. Так, наприклад, для завдань аналізу трафіку та структури мережі великої компанії незалежно від галузевої спрямованості (банки, торгівля, логістика, промисловість, сільське господарство та ін.), групи об'єктів з високим ступенем подібності, що формуються в результаті кластеризації, можуть включати такі категорії [5–8, 11–20, 23, 24]:

– типи трафіку, наприклад, веб-трафік (HTTP/HTTPS запити та відповіді, відвідування веб-сайтів, використання веб-додатків); поштовий трафік (SMTP, IMAP, POP3 та інші протоколи електронної пошти); файловий трафік (передача даних через FTP, SFTP, SMB та інші протоколи обміну файлами); потоковий трафік (відео та аудіо потоки, що використовують протоколи, такі як RTP, RTSP, HLS);

– активність користувача, наприклад, групи користувачів (співробітники певних відділів, віддалені користувачі, адміністративний персонал);

– поведінкові патерни, наприклад, типові часові рамки активності, звички використання додатків і сервісів.

– мережеві пристрої, наприклад, типи пристроїв (робочі станції, сервери, мобільні пристрої, IoT-пристрої); функціональні групи (пристрої з однаковими функціями, такі як сервери баз даних, веб-сервери, мережеві шлюзи);

– застосунки та сервіси, наприклад, до цієї категорії можна віднести – кластери додатків (угруповання за використовуваними протоколами та типами даних, що передаються додатками); сервіси (веб-сервіси, бази даних, хмарні сервіси, служби зберігання даних);

– аномалії та загрози, наприклад, аномальна поведінка (трафік, який відхиляється від типових патернів, що може сигналізувати про можливі атаки чи зловживання) тощо.

Подібні та інші групи, які не увійшли до вищенаведеного переліку, допомагають у подальшому аналізі для оптимізації трафіку, виявлення аномалій, покращення безпеки та підвищення ефективності мережі. Таким чином, кластеризація дозволяє ідентифікувати та ізолювати різні типи трафіку та їх джерела, що полегшує управління мережею та забезпечення її надійної роботи.

Постановка проблеми. Проблема полягає в необхідності ефективної обробки та інтерпретації великих обсягів інформації, що отримується з різних джерел, таких як серверні логи, датчики трафіку та геолокаційні дані. Без впровадження сучасних рішень для обробки та аналізу даних, компанії ризикують втратити конкурентні переваги через низьку швидкість аналізу, високі витрати на обробку та нестачу інформації для прийняття обґрунтованих рішень. Необхідність розробки адаптивних та масштабованих систем для обробки та аналізу трафіку стає актуальною, враховуючи динамічні зміни в бізнес-середовищі та зростаючі вимоги до оперативності та точності даних.

Аналіз останніх досліджень та публікацій. Критерієм якості кластеризації є функціонал, який залежить від об'єктів усередині груп та відстаней між ними [2–05, 8, 13, 14, 17, 18, 20, 24]. На відміну від класифікації, при кластеризації спочатку не визначено число та властивості класів (кластерів), що дає можливість, наприклад, для нашого завдання адаптивно аналізувати трафік та виявляти нові аномалії та тенденції.

Відповідно до [14, 17, 18], особливості кластеризації включають:

– можливість виявлення раніше невідомих класів об'єктів з урахуванням початкових характеристик;

– здатність ефективно обробляти великі обсяги даних за короткий термін.

Для підвищення стійкості рішень у задачах кластеризації можна використовувати ансамблі алгоритмів, які формують колективне рішення з урахуванням думок всіх учасників ансамблю. Цей підхід є особливо актуальним при аналізі трафіку, де обсяг даних великий, а структура трафіку може бути складною та різноманітною.

Таким чином, основні переваги та особливості кластерного аналізу роблять його незамінним інструментом для оптимізації та масштабування мережевого трафіку у великих компаніях. Тому в роботі основна увага приділяється розробці ансамблю алгоритмів кластеризації на основі динамічних метрик відстаней для аналізу великих обсягів трафіку, що дозволяє значно покращити якість та швидкість аналізу.

Результати дослідження. У сучасному світі обсяги даних, що генеруються корпоративними мережами, зростають з неймовірною швидкістю, що створює нові виклики та можливості для аналізу мережевого трафіку з метою його оптимізації та підвищення безпеки, пропускної спроможності та ін. Одним з ефективних методів обробки великих обсягів мережевого трафіку є кластеризація даних [1–8, 11– 8, 20, 24]. Кластеризація дозволяє сегментувати трафік на однорідні групи, що сприяє більш точному виявленню закономірностей, аномалій та потенційних загроз.

Дослідження останніх років, що розглядаються далі, демонструють значний прогрес у використанні методів кластерного аналізу (КА) для вирішення завдань, пов'язаних із аналізом мережевого трафіку. Зокрема, вчені у своїх роботах [8, 13, 14, 17, 18] розглядають різні алгоритми кластеризації та їх модифікації для підвищення точності та швидкості обробки даних. Літературний аналіз [6–10, 15–19, 23] показав, що поєднання кластеризації з методами машинного навчання (МН) та штучного інтелекту (ШІ) також є перспективним напрямом.

У даному дослідженні короткий аналіз подано ключові роботи, в основному за останні 5–10 років, присвячені застосуванню КА для дослідження трафіку компаній, їх основні висновки отримані в ході цих досліджень і застосовані авторами розглянутих робіт і методології. Такий аналіз дозволить краще зрозуміти поточний стан досліджень у цій галузі та визначити напрями для подальшого розвитку нових досліджень у даній сфері.

У [20] розглядаються дванадцять існуючих методів кластеризації. Огляд, виконаний авторами роботи, дозволив розглянути існуючі проблеми та рекомендації для подальших досліджень у галузі кластеризації потоків трафіку.

У [13] запропоновано алгоритм класифікації трафіку на основі поліпшеної кластеризації K-means. Авторами дослідження наочно продемонстровано принцип роботи алгоритму для цієї задачі, а також проведено порівняння отриманих результатів та верифікацію на тестовому наборі даних.

У [17] автори порівнюють алгоритми EM, DBScan та RAIN для аналізу трафіку корпоративних мереж.

Зауважимо, що ідея використання кластеризації даних для аналізу мережевого трафіку не є новою. Її коріння сягає кінця 60-х, початок 70-х років минулого століття, коли почалися перші спроби систематизувати і структурувати великі обсяги даних, для виявлення закономірностей і аномалій у мережевих взаємодіях. З того часу дослідники постійно вдосконалюють методи кластеризації, адаптуючи їх до нових викликів, пов'язаних із зростанням обсягу та складності мережевого трафіку.

Вже наприкінці 60-х і на початку 70-х років минулого століття, з'явилися перші публікації, що розглядаються нижче, присвячені кластерному аналізу мережевих даних. Ці ранні роботи, наприклад, [14] заклали основу для подальших досліджень та розробок. З того часу методи кластеризації розвивалися в різних напрямках, від простих алгоритмів, таких як K-means, до складних ансамблевих методів, що поєднують переваги декількох алгоритмів для досягнення більш точних і надійних результатів.

Ранні дослідження [14] показали, що кластеризація може бути ефективно використана для сегментації мережевого трафіку, дозволяючи виділити однорідні групи даних та проводити їх детальніший аналіз.

В останні десятиліття, з розвитком технологій Big Data та збільшенням обчислювальних потужностей, інтерес до КА даних у контексті вивчення та оптимізації мережевого трафіку значно зріс. Сучасні

дослідження фокусуються на розробці більш складних та точних методів, включаючи використання МН та ШІ для покращення якості кластеризації.

У [14] запропоновано неієрархічний метод розбиття. У [18] метод K-means був застосований для розбиття наборів даних на кластери на основі заздалегідь визначеної кількості спочатку вибраних центроїдів (k). Згідно з висновками автора [18] удосконалений алгоритм, запропонований у даній роботі, дозволяє використовувати евклідову відстань для зменшення помилок, що виникають при обчисленні середніх квадратів з цільової функції.

У роботі [8] запропоновано підхід до двоетапної класифікації мережевого трафіку з використанням кластеризації K-means для покращення управління якістю обслуговування (QoS). Метою даного дослідження було розробити ефективний класифікатор, здатний розпізнавати цільові програми та виявляти невідомі потоки (шум) у мережі. Запропонований метод класифікації ґрунтувався на аналізі поведінки потоків трафіку та складався з двох основних фаз: 1) фаза присвоєння; 2) фаза маркування. У фазі присвоєння потоки призначаються певному кластеру, а фазі маркування використовується алгоритм присвоєння потокам відповідних міток. Ці етапи дозволяють оновлювати класифікатор для його подальшого використання у системі управління трафіком.

У роботі [24] автори розглядають метод BIRCH, який включає масштабованість у модель кластеризації. У своєму дослідженні вони використовують дерево ознак кластеризації (CF-дерево), та багаторівневу кластеризацію для обробки великих наборів даних через два основні етапи, кожен з яких має додаткову фазу. На першому етапі великі набори даних або об'єкти даних стискаються в компактне CF-дерево, що зберігає базову структуру кластерів. На другому етапі застосовується агломераційний алгоритм у поєднанні з іншими гнучкими методами кластеризації для створення вихідних кластерів, які потім уточнюються на основі їх центроїдів.

У роботі [12] автори запропонували ієрархічний кластеризаційний алгоритм під назвою Clustering Using Representatives (CURE). Даний алгоритм був розроблений для кластеризації великих наборів даних і здатний ефективно справлятися зі спотвореннями, спричиненими викидами. Він особливо добре підходить для кластерів довільної та несферичної форми з високою дисперсією. На відміну від алгоритму BIRCH, CURE спочатку випадково вибирає підвибірку даних і поділяє її на секції перед початком кластеризації. Потім ці секції частково групуються для видалення викидів. Після видалення викидів часткові кластери повторно кластеризуються для отримання дрібніших кластерів, які потім поєднуються в остаточні кластери. Такий підхід дозволяє покращити стійкість алгоритму до викидів та точніше виявляти структуру даних.

У роботі [11] автори запропонували алгоритм кластеризації, заснований на щільності, названий DBSCAN. Метою цього алгоритму було виявлення кластерів довільної форми та визначення шуму в даних. Грунтуючись на щільності, DBSCAN враховує якість кластерів та їхню здатність ідентифікувати шумові точки. Для визначення кластерів у DBSCAN використовуються два основні параметри: Eps та MinPts. Параметр Eps визначає радіус окружності точки (P), який визначає досяжність щільності, а MinPts – мінімальна кількість точок на окружності Eps, необхідне для формування кластера. Процес кластеризації починається з довільної точки (a), і якщо відстань від точки (a) до (P) менше або дорівнює Eps, точка додається в кластер. Цей процес триває ітеративно для включення нових точок у кластер. DBSCAN має чутливість до вибору параметрів Eps та MinPts, що може ускладнювати їх налаштування. Однак, при правильному налаштуванні, DBSCAN ефективно виявляє кластери довільної форми та стійкий до шуму даних.

У роботі [6] автори запропонували алгоритм OPTICS для подолання недоліків, властивих DBSCAN. На відміну від DBSCAN, OPTICS менш чутливий до налаштування параметрів. Як і інші методи, що ґрунтуються на щільності, OPTICS генерує порядок кластеризації, який містить інформацію про структуру кластерів для широкого діапазону значень параметрів. OPTICS добре масштабується при зміні значень Eps (ε) в діапазоні від 10 000 до 100 000. Це дозволяє алгоритму працювати швидко та ефективно навіть з великими обсягами даних, притаманних трафіку компаній, коли обсяг даних, для короткого проміжку часу, може досягати кількох десятків терабайт. Таким чином, OPTICS забезпечує, на думку авторів, більш гнучку та детальну ідентифікацію кластерних структур порівняно з DBSCAN.

У роботі [19] автори застосували гібридне рішення на основі алгоритмів OPTICS та DBSCAN для вирішення проблеми вибору відповідного порога щільності при виявленні спільнот у соціальних мережах. Вибір правильного порога щільності сприяє отриманню змістовних кластерів. Оскільки щільність визначається функцією відстані, використання OPTICS дозволило авторам вибрати оптимальне значення параметра Eps для DBSCAN, а також реалізувати результати використання змінних порогових значень щільності. Питання про те, чи можливе справжнє визначення спільноти в соціальних мережах, залишається відкритим, як показав аналіз авторів, які проводили це дослідження.

Дослідження в галузі класифікації IP та трафіку з використанням унікальних характеристик потоку також виявилися дуже ефективними. У роботі [23] автори запропонували автоматизований метод класифікації, що ґрунтується на статистичних характеристиках потоку, з використанням NetMate [16]. Цей неконтрольований метод використовує алгоритм максимізації очікувань [15] та алгоритм AutoClass [8]. Пакети спочатку розбиваються на двоспрямовані потоки обчислення характеристик потоку. Разом з атрибутами моделі потоків класи можуть бути вивчені для подальшої класифікації нових потоків. Результати можуть бути використані для оцінки та інших цілей QoS.

Напівконтрольовані методи (semi-supervised clustering) також призвели до нового виміру досліджень у галузі КА. Наприклад, у роботі [10] авторами представлені результати досліджень, присвячених кластеризації з використанням контрольованих та неконтрольованих методів. У проектуванні використовувалися контрольні точки пакетів. Автори досліджували класифікацію трафіку з використанням характеристик потоку в додатках та запропонували напівконтрольований метод класифікації трафіку з відомих та невідомих додатків. Класифікатор навчається шляхом порівняння потоків трафіку переважно з потоками без міток, при цьому мінімально включаються потоки з мітками.

Щоб підвищити точність методів класифікації, у роботі [21] автори запропонували напівконтрольовану стратегію, яка називається обмеженими K-means на основі множин. Статистичні характеристики потоку вилучаються разом із деякою довідковою інформацією про потоки TCP/IP. Для моделювання даних, що спостерігаються використовується гауссова суміш щільностей. У роботі було встановлено, що введення дискретних ознак в кластеризацію потоків може підвищити точність кластеризації. На основі ступеня подібності або відмінності функцій потоку, вони групуються відповідно до п'яти міток кортежів, що включають вихідні та цільові IP-адреси, вихідні та цільові порти, а також протокол, що використовується портом. Потоки, що мають схожість у різних застосунках, швидше за все, будуть згруповані у певний кластер.

У рамках програмно-конфігурованих мереж (SDN) [22] автори розробили метод класифікації трафіку, поєднуючи вимоги щодо якості обслуговування з реалізацією Deep Packet Inspection (DPI). Вони виявляли вхідні потоки з тривалим терміном служби за допомогою комутатора SDN. Використовуючи значення пакета Херста, порту та середнього часу між прибуттям пакетів як вхідні дані у функцію співставлення, трафік класифікувався за відповідними класами QoS. Статистичні ознаки було зібрано, і черги класів формувалися з потоків. Потім потоки класифікувалися за відповідними класами QoS.

Для дослідження якості обслуговування, застосовуючи генеративну модель (прихована марковська модель, ПММ) для напівконтрольованого навчання послідовностей [14] автори запропонували новий метод класифікації трафіку на рівні пакетів. Використання послідовності ПММ кваліфікує цей підхід як напівсупервізійний. Грунтуючись на характеристиках розміру та часу між пакетами, автори розробили класифікацію, що спирається на агреговані характеристики реального мережевого трафіку. Даний метод виявився придатним для використання в зашифрованому трафіку.

В [7] досліджено можливість аналізу мережевого трафіку з використанням методів машинного навчання (ММН). Для зменшення розмірності даних було застосовано вибірка найбільш значущих ознак (15 з 87 ознак) на реальному наборі даних із понад 3 мільйонів екземплярів. Потім була застосована кластеризація K-means для кращого розуміння та розрізнення поведінки трафіку. Результати показали хорошу кореляцію між екземплярами в одному кластері, отриманому за допомогою навчання без учителя.

На підставі виконаного огляду попередніх досліджень, була сформована таблиця 1, в якій узагальнено отримані результати аналізу.

Таким чином, як показав, виконаний аналіз публікацій, для розробки ефективних методів аналізу та обробки великих даних (Big Data), на основі кластерних колективних рішень для аналізу трафіку у великих компаніях, дослідники виявляють великий інтерес до створення більш точних методів класифікації та визначенню моделей трафіку в реальному часі у мережевій безпеці та інших мережевих рішеннях. Багато моделей було сформульовано на основі існуючих неконтрольованих і напівконтрольованих методів кластеризації. Ці моделі включають методи, що демонструють здатність алгоритмів справлятися з шумом, а також їхню продуктивність і здатність класифікувати великі набори даних мережевого трафіку в реальному часі. Хоча класичний підхід K-means є основою розробки кількох методів напівконтрольованої кластеризації, пов’язана з ним обчислювальна складність, обмежує його застосування за умов обмежених обчислювальних ресурсів. Однак, наскільки нам відомо, існує обмежена кількість досліджень, присвячених аналізу роботи алгоритмів за певних параметрів QoS, що є метою для подальшого вивчення.

Таблиця 1

Порівняльний аналіз методів кластеризації для аналізу трафіку у великих компаніях (складено авторами на підставі аналізу літературних джерел [1–24])

Автори та джерело	Цілі	Використовуваний метод кластеризації	Параметри кластеризації	Обмеження	Результат
Lloyd, S. [18]	Зменшити вплив шумів при обчисленні середніх квадратів центроїдів у процесі формування кластерів	Класичний K-means	Функція відстані	Чутливість до шумів даних, не якісна кластеризація при поганій ініціалізації	Формує щільно пов'язані кластери в порівнянні з традиційними ієрархічними методами
Zhang, T. та ін. [24]	Підвищити ефективність використання ресурсів для обробки великих наборів даних	Алгоритм BIRCH (ієрархічний)	Дерево ознак (CF-дерево)	Чутливість до вставки даних (шум), більш високе робоче навантаження на процесор	Обробляє великі набори даних за менший час порівняно з K-Means
Guba, S. та ін. [12]	Дозволяє ідентифікувати несферичні кластери довільної форми та протистояти викидам у великих наборах даних	Алгоритм CURE (ієрархічний)	Репрезентативні точки для кластерів, коефіцієнт стиснення	Висока обчислювальна складність, а отже, і висока вартість	Формує кластери високої якості, час виконання на 50% менший у порівнянні з BIRCH [24] зі збільшенням кількості точок
Ester, M. та ін. [11]	Підвищити якість кластерів за допомогою можливостей алгоритму ідентифікації шумів	Алгоритм DBSCAN	Досяжність (Eps), максимальний радіус сусідства (MinPts)	Чутливість до параметрів (Eps і MinPts), складність обчислення параметрів, час виконання збільшується зі зростанням бази даних	Метод здатний ідентифікувати та виявляти прояви шуму. Також час виконання кращий, ніж алгоритм CLARANS
Ankerst, M. та ін. [6]	Подолати обмеження DBSCAN [14], пов'язані з чутливістю до параметрів	Алгоритм OPTICS (на основі розподілу густини)	Досяжність (Epx), максимальний радіус сусідства (MinPts)	Складність керування параметрами при кластеризації зі зростаючим порядком, час виконання можна порівняти з DBSCAN [14], але з нижчими налаштуваннями параметрів	Діаграма досяжності нечутлива до вхідних даних кластеризації в порівнянні з DBSCAN та іншими алгоритмами, час виконання можна порівняти з DBSCAN, але з нижчими налаштуваннями початкових параметрів
Subramani, K. та ін. [19]	Вибрати відповідний поріг щільності для виявлення спільнот у соціальних мережах	Гібридний підхід (OPTICS в DBSCAN)	поріг щільності	Обчислювальна складність гібридного підходу не обговорюється, визначення порогу щільності може призвести до раптової зміни та залежить від припущень програми	Гібридний підхід забезпечує чітке розуміння кластеризації, простоту вибору порога щільності за допомогою запропонованого методу
Zander, S. та ін. [23]	Підвищити загальну внутрішньокласову однорідність	Алгоритм ймовірнісного підходу (Expectation Maximization and Mixture Models (AutoClass))	Статистичні характеристики (внутрішньо-класова однорідність як метрика)	Продуктивність на великих наборах даних із зростаючою кількістю класів	Досягає середньої точності 85% при кластеризації

Wang, Y. та ін. [21]	Розробка методу класифікації інтернет-трафіку з використанням обмеженої кластеризації.	Обмежені K-meansна основі множин	Статистичні характеристики потоку, інформація про потоки TCP/IP, гаусова щільність суміші, дискретні ознаки	Можливість обробки великого обсягу даних та високі обчислювальні витрати	Поліпшення точності кластеризації, особливо для потоків зі схожими характеристиками у різних додатках
Wang, P. та ін. [22]	Класифікація мережевого трафіку з урахуванням вимог щодо якості обслуговування (QoS) у SDN.	Напівсупервізійне навчання з використанням Deep Packet Inspection (DPI)	Параметри DPI, значення пакета Херста, порти, середній час між прибуттям пакетів	Залежать від точності та доступності даних для навчання; висока обчислювальна складність	Ефективна класифікація трафіку за відповідними класами QoS, покращення управління мережею
Casas, P. та ін. [7]	Розробка платформи для аналізу мережевого трафіку з використанням технологій Big Data	Не зазначено конкретно, використовуються методи в області Big Data Analytics	Параметри аналізу включають часові ряди, статистичні характеристики трафіку.	Складність інтеграції та обробки великих обсягів даних, необхідність значних обчислювальних ресурсів	Ефективна платформа для моніторингу та аналізу мережевого трафіку з можливістю обробки великих обсягів даних у реальному часі

Резюмуючи виконаний огляд попередніх досліджень, можна зробити висновок – завдання розробки системи аналізу трафіку для великої компанії з метою оптимізації маршрутизації, зниження витрат та підвищення швидкості доставки може бути вирішена за допомогою створення кластерної платформи для розподіленої обробки даних з використанням сховищ даних певного типу. Для збирання та зберігання даних можна, наприклад, використовувати логи серверів, див. рис. 1. дані від датчиків трафіку, див. рис. 2., інформацію про розташування користувачів, див. рис. 3. та дані про маршрути, див. рис. 4., 5. і іншу інформацію, специфічну для кожного з бізнес-процесів, які будуть розглянуті в наступних публікаціях.

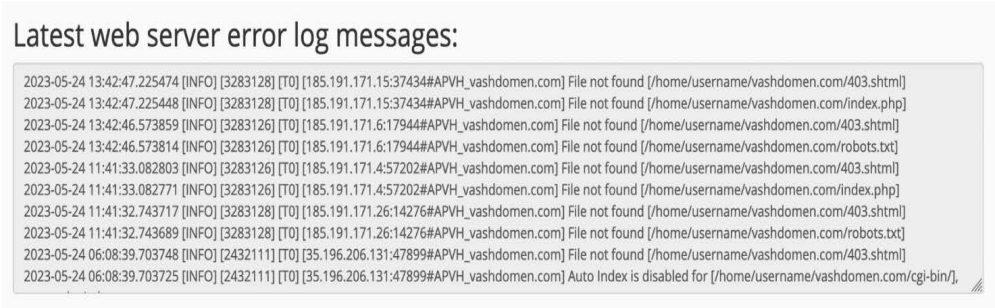

Рис. 1. Приклад логів сервера

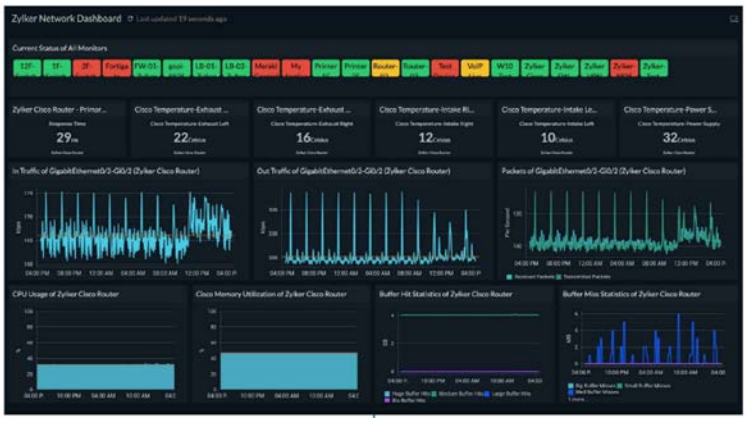

Рис. 2. Приклад даних від датчиків трафіку

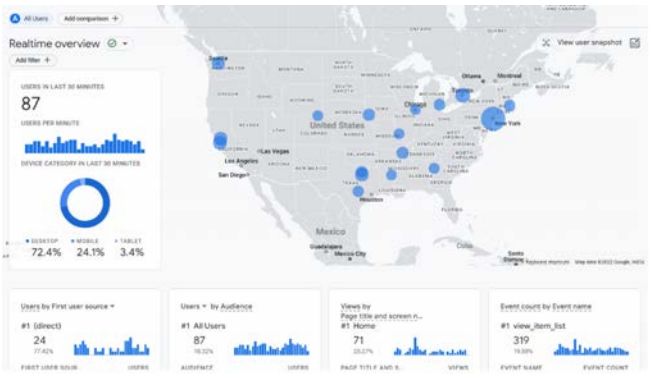

Рис. 3. Приклад інформації про розташування користувачів

Передобробка даних, згідно з класичним підходом, включатиме їх очищення, нормалізацію та перетворення в єдиний формат для структурованого зберігання в кластерній системі, що забезпечує швидкий доступ до потрібних даних.

Аналіз даних можна здійснювати, знов-таки виходячи зі специфіки бізнес-процесів, за допомогою алгоритмів машинного навчання (МН) для класифікації, регресії та кластеризації, аналізу часових рядів для прогнозування трафіку, просторового аналізу для оптимізації маршрутів доставки та аналізу мережевої топології для виявлення вузьких місць та оптимізації маршрутизації. Візуалізацію результатів можна виконати через інтерактивні панелі управління, графіки та діаграми, що дозволить представляти ключові показники та аналітичні дані у наочній формі. Інструменти прийняття рішень нададуть інформацію, необхідну для оптимізації трафіку.

```
Gateway of last resort is not set

R    17.0.0.0/8 [120/2] via 19.1.1.2, 00:00:12, Serial0/1
                [120/2] via 12.1.1.2, 00:00:10, Serial0/2
R    16.0.0.0/8 [120/1] via 19.1.1.2, 00:00:12, Serial0/1
                [120/1] via 12.1.1.2, 00:00:10, Serial0/2
C    19.0.0.0/8 is directly connected, Serial0/1
R    18.0.0.0/8 [120/1] via 19.1.1.2, 00:00:12, Serial0/1
R    192.168.8.0/24 [120/4] via 19.1.1.2, 00:00:12, Serial0/1
                    [120/4] via 12.1.1.2, 00:00:10, Serial0/2
R    21.0.0.0/8 [120/3] via 19.1.1.2, 00:00:15, Serial0/1
                [120/3] via 12.1.1.2, 00:00:13, Serial0/2
R    192.168.9.0/24 [120/4] via 19.1.1.2, 00:00:15, Serial0/1
                    [120/4] via 12.1.1.2, 00:00:13, Serial0/2
R    20.0.0.0/8 [120/4] via 19.1.1.2, 00:00:15, Serial0/1
                [120/4] via 12.1.1.2, 00:00:13, Serial0/2
R    192.168.4.0/24 [120/2] via 12.1.1.2, 00:00:17, Serial0/2
                    [120/2] via 10.1.1.2, 00:00:12, Serial0/0
R    192.168.5.0/24 [120/2] via 19.1.1.2, 00:00:19, Serial0/1
                    [120/2] via 12.1.1.2, 00:00:17, Serial0/2
C    10.0.0.0/8 is directly connected, Serial0/0
R    192.168.6.0/24 [120/1] via 19.1.1.2, 00:00:20, Serial0/1
R    11.0.0.0/8 [120/1] via 12.1.1.2, 00:00:18, Serial0/2
                [120/1] via 10.1.1.2, 00:00:13, Serial0/0
R    192.168.7.0/24 [120/3] via 19.1.1.2, 00:00:20, Serial0/1
                    [120/3] via 12.1.1.2, 00:00:18, Serial0/2
C    12.0.0.0/8 is directly connected, Serial0/2
C    192.168.1.0/24 is directly connected, FastEthernet0/1
R    13.0.0.0/8 [120/1] via 10.1.1.2, 00:00:14, Serial0/0
R    192.168.2.0/24 [120/1] via 10.1.1.2, 00:00:15, Serial0/0
R    14.0.0.0/8 [120/1] via 12.1.1.2, 00:00:20, Serial0/2
R    192.168.3.0/24 [120/1] via 12.1.1.2, 00:00:20, Serial0/2
R    15.0.0.0/8 [120/1] via 12.1.1.2, 00:00:20, Serial0/2
```

Рис. 4. Таблиця маршрутизації маршрутизатора R1

Рис. 5. Топологія мережі

Для ілюстрації такого підходу наведемо невеликий приклад для популярних інтернет-магазинів. Даний приклад реалізації буде включати аналіз трафіку інтернет-магазину, де дані збираються з логів серверів, інформації про замовлення та місцезнаходження користувачів. Обробка даних дозволить агрегувати їх за часом, місцем розташування та типом замовлень. Алгоритми кластеризації виявить групи користувачів зі схожою поведінкою, а застосування регресії дозволить спрогнозувати попит на певні товари. На заключному етапі, можна візуалізувати результати у вигляді теплових карток завантаженості серверів та графіків попиту по регіонах, що допоможе оптимізувати розподіл ресурсів та доставку товарів споживачам. Це типовий приклад, а конкретні рішення з різних галузей економіки ми розглянемо у наступному параграфі даного розділу роботи.

Таким чином, виходячи з результатів виконаного огляду та аналізу попередніх досліджень можна констатувати, що кластерні рішення пропонують масштабованість, високу швидкість виконання завдань завдяки розподіленій обробці, доступність даних та сервісів, а також гнучкість у використанні різних інструментів та технологій. Однак реалізація таких систем все ще пов'язана з технічними складнощами, що вимагають знань у галузі Big Data, ММН та кластерних систем, а також із витратами на обладнання, програмне забезпечення та кваліфікованих фахівців. Однак, незважаючи на вищезазначені складності, використання кластерних рішень для аналізу Big Data потенційно може надати компаніям конкурентні переваги за рахунок оптимізації процесів, прийняття більш обґрунтованих рішень та підвищення ефективності роботи.

Висновки. У процесі досліджень було отримано такі основні результати.

Виконано огляд попередніх досліджень у завданнях, пов'язаних із дослідженням трафіку компаній на основі кластеризації даних. Встановлено, що завдання розробки системи аналізу трафіку для великої компанії з метою оптимізації маршрутизації, зниження витрат та підвищення швидкості доставки може бути вирішено за допомогою створення кластерної платформи для розподіленої обробки даних з використанням сховищ даних різного типу.

Продемонстровано, що для збирання та зберігання даних можна, наприклад, використовувати логи серверів, дані від датчиків трафіку, інформацію про місцезнаходження користувачів та дані про маршрути тощо.

Встановлено, що кластерні рішення пропонують масштабованість, високу швидкість виконання завдань, завдяки розподіленій обробці, доступність даних та сервісів, а також гнучкість у використанні різних інструментів та технологій. Однак реалізація таких систем все ще пов'язана з технічними складнощами, що вимагають знань у галузі Big Data, методах машинного навчання та кластерних систем, а також витратами на обладнання, програмне забезпечення (ПЗ) та кваліфікованих фахівців. Однак, незважаючи на вищезазначені складнощі, використання кластерних рішень для аналізу Big Data потенційно може надати компаніям конкурентні переваги за рахунок оптимізації процесів, прийняття більш обґрунтованих рішень та підвищення ефективності роботи.

Список використаних джерел:

1. Джулій В. М., Солодєєва Л. В., Мірошніченко О. В. Метод класифікації додатків трафіка комп'ютерних мереж на основі машинного навчання в умовах невизначеності. *Наукові праці*. 2022. С. 73–82. DOI: https://doi.org/10.17721/2519-481X/2022/74-07.

2. Лунгол О. Огляд методів та стратегій кібербезпеки засобами штучного інтелекту. Кібербезпека: освіта, наука, техніка. 2024. Т. 1(25). С. 379–389.

3. Мамарєв В. М. Аналіз сучасних методів виявлення атак на ресурси інформаційно-телекомунікаційних систем. *Ukrainian Information Security Research Journal*. 2011. Т. 13(2 (51)).

4. Морозов Б. Дослідження методів аналізу мережевого трафіку. Матеріали IX Всеукраїнської студентської науково-технічної конференції „Природничі та гуманітарні науки. Актуальні питання". 2016. Т. 1. С. 91–92.

5. Рубан І. В., Мартовицький В. О., Партика С. О. Класифікація методів виявлення аномалій в інформаційних системах. *Системи озброєння і військова техніка*. 2016. Т. 3. С. 100–105.

6. Ankerst M., Breunig M. M., Kriegel H. P., Sander J. OPTICS: Ordering points to identify the clustering structure. *ACM Sigmod record*. 1999. 28(2), pp. 49–60.

7. Aouedi O., Piamrat K., Hamma S., Perera J.M. Network traffic analysis using machine learning: an unsupervised approach to understand and slice your network. *Annals of Telecommunications*. 2022. T. 77(5). pp. 297–309.

8. Cheeseman P. C., Stutz J. C. Bayesian classification (AutoClass): theory and results. Advances in knowledge discovery and data mining. 1996. N. 180. pp. 153–180.

9. Dainotti A., De Donato W., Pescape A., Rossi P.S. Classification of network traffic via packet-level hidden markov models. IEEE GLOBECOM 2008-2008 IEEE Global Telecommunications Conference. 2008. pp. 1–5.

10. Erman J., Mahanti A., Arlitt M., Cohen I., Williamson C. Offline/realtime traffic classification using semi-supervised learning. Performance Evaluation. 2007. N 64, pp. 1194–1213.

11. Ester M., Kriegel H.P., Sander J., Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD*. Vol. 96, No. 34. 1996. pp. 226–231.

12. Guha S., Rastogi R., Shim K. CURE: An efficient clustering algorithm for large databases. *ACM Sigmod record*. 1998. T. 27(2). pp. 73–84.

13. Li J., Zhang H., Tang D., Lin C. Traffic classification using cluster analysis. 2021 International Conference on Computer Information Science and Artificial Intelligence (CISAI). 2021. pp. 463–467.

14. MacQueen J. Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. 1967. T. 1, No. 14. pp. 281–297.

15. McGregor A., Hall M., Lorier P., Brunskill J. Flow clustering using machine learning techniques. Passive and Active Network Measurement: 5th International Workshop, PAM 2004. 2004. pp. 205–214.

16. NetMate Meter. URL: http://sourceforge.net/projects/netmate-meter.

17. Rodriguez Rodriguez J. E., Garcia V.H.M., Usaquén M.A.O. Corporate networks traffic analysis for knowledge management based on random interactions clustering algorithm. Knowledge Management in Organizations: 13th International Conference, KMO 2018. 2018. pp. 523–536.

18. S. Lloyd, "Least squares quantization in PCM", IEEE transactions on information theory. 1982. vol. 28, no. 2, pp. 129–137.

19. Subramani K., Velkov A., Ntoutsi I., Kroger P., Kriegel H. P. Density-based community detection in social networks. In 2011 IEEE 5th International Conference on Internet Multimedia Systems Architecture and Application. 2011. pp. 1–8.

20. Takyi K., Bagga A., Goopta P. Clustering techniques for traffic classification: A comprehensive review. 2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO). 2018. pp. 224–230.

21. Wang Y., Xiang Y., Zhang J., Zhou W., Wei G., Yang L.T. Internet traffic classification using constrained clustering. IEEE transactions on parallel and distributed systems. 2013. T. 25(11). pp. 2932–2943.

22. Wang P., Lin S.C., Luo M. A framework for QoS-aware traffic classification using semi-supervised machine learning in SDNs. 2016 IEEE international conference on services computing (SCC). 2016. pp. 760–765.

23. Zander S., Nguyen T., Armitage G. Automated traffic classification and application identification using machine learning. The IEEE Conference on Local Computer Networks 30th Anniversary (LCN'05). 2005. pp. 250–257.

24. Zhang Tian, Raghu Ramakrishnan, Miron Livny. BIRCH: A new data clustering algorithm and its applications. Data mining and knowledge discovery. 1997. N. 1. pp. 141–182.