МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Харківський національний університет радіоелектроніки
Факультет Комп'ютерних наук
Кафедра Програмної інженерії

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

_____другий (магістерський)_____
(рівень вищої освіти)

«Дослідження методів кластеризації даних при аналізі спільнот у соціальних мережах»

Виконав:
студент __2__ курсу групи ____ІПЗм-21-3__
_____Циба І.О._____
(прізвище, ініціали)

Спеціальність 121 – Інженерія програмного

_____забезпечення_____

Тип програми _____освітньо-наукова_____

Керівник _____доцент Кобзєв В.Г._____
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ З.В. Дудар

2023 р.

Харківський національний університет радіоелектроніки

Факультет _____Комп'ютерних наук_____
(повна назва)

Кафедра _____Програмної інженерії_____
(повна назва)

Рівень вищої освіти_____другий (магістерський)_____

Спеціальність_____121 – Інженерія програмного забезпечення_____
(код і повна назва спеціальності)

Тип програми _____освітньо-наукова програма _____
(освітньо-професійна або освітньо-наукова)

Освітня програма_____Інженерія програмного забезпечення _____
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«_____» _____20 ____ р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

студента _____*Циби Івана Олександровича*_____
(прізвище, ім'я, по батькові)

1. Тема роботи «Дослідження ефективності застосування методів кластеризації при аналізі спільнот у соціальних мережах»

затверджена наказом університету від «29» березня 2023 р. № 302Ст

2. Термін подання роботи до екзаменаційної комісії «24» травня 2023р.

3. Вихідні дані до роботи *методи та алгоритми кластеризації, бібліотеки Python, аналіз алгоритмів кластеризації, метрики кластеризації.*

4. Перелік питань, що потрібно опрацювати в роботі *вступ, аналіз предметної області, аналіз алгоритмів кластеризації, критерії їх оцінки, постановка задачі, проведення експериментів та аналіз їх результатів, формування подальших рекомендацій.*_____

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Аналіз предметної галузі	26.03.2023	виконано
2	Виявлення проблематики галузі	02.04.2023	виконано
3	Ознайомлення зі структурою роботи та отримання індивідуального завдання	02.04.2023	виконано
4	Дослідження матеріалів та наукових робіт	09.04.2023	виконано
5	Постановка задачі	16.04.2023	виконано
6	Планування експериментальної частини дослідження	16.04.2023	виконано
7	Підготовка пояснювальної записки	16.04.2023	виконано
8	Підготовка презентації та доповіді	23.04.2023	виконано
6	Попередній захист	23.04.2023	виконано
7	Перевірка на академічний плагіат	22.05.2023	виконано
8	Нормоконтроль	22.05.2023	виконано
9	Рецензування	22.05.2023	виконано
10	Знесення диплома в електронний архів	22.05.2023	виконано
11	Допуск до захисту у зав. кафедри	23.05.2023	виконано

Дата видачі завдання 03.04.2023 р.

Студент _____

(підпис)

Керівник роботи_____ доцент Кобзєв В.Г.

(підпис)

РЕФЕРАТ

Пояснювальна записка кваліфікаційної роботи: 73 pages, 10 figures, 11 tables, 8 formulas, 24 sources.

АЛГОРИТМИ, АНАЛІЗ ДАНИХ, ВИЯВЛЕННЯ СПІЛЬНОТ, ГРАФИ, КЛАСТЕРИЗАЦІЯ, МЕТРИКИ УСПІШНОСТІ, МОДУЛЯРНІСТЬ, СПІЛЬНОТИ, СОЦІАЛЬНІ МЕРЕЖІ, СТРУКТУРА МЕРЕЖІ.

Об’єкт дослідження – спільноти у соціальних мережах.

Предмет дослідження – методи кластеризації даних для аналізу спільнот у соціальних мережах.

Метою роботи є дослідити та оцінити різні методи кластеризації, вибрати найефективніші для аналізу спільнот у соціальних мережах, та застосувати їх для розробки практичного дослідження

Методи розробки – аналіз наукової літератури, порівняння методів кластеризації, вибір атрибутів, підготовка даних, визначення метрик успішності, проведення практичних експериментів.

Об’єктом розробки є алгоритми та програмне забезпечення для кластеризації та аналізу спільнот у соціальних мережах на основі вибраних методів.

CLUSTERING, COMMUNITIES, SOCIAL NETWORKS, DATA ANALYSIS, ALGORITHMS, GRAPHS, NETWORK STRUCTURE, MODULARITY, COMMUNITY DETECTION, PERFORMANCE METRICS

The object of research is a social dimension.

The subject of the research is the methods of data clustering for the analysis of skills in social networks.

The method of work is to evaluate and evaluate the different methods of clustering, to choose the most effective for the analysis of the characteristics of social networks, and to assess them for the development of a practical follow-up

Research methods – analysis of scientific literature, matching of clustering methods, selection of attributes, preparation of data, designation of success metrics, practical experiments.

The object of research is algorithms and software for clustering and analysis of social networks on the basis of selected methods.

Я, Циба Іван Олександрович, студент групи ІПЗм-21-3, здобувач вищої освіти на другому (магістерському) рівні, кафедра Програмної інженерії, заявляю: моя кваліфікаційна робота на тему «Дослідження ефективності застосування методів кластеризації при аналізі спільнот у соціальних мережах», що буде представлена до ЕК для публічного захисту, виконана самостійно, в ній не містяться елементи плагіату і вона може бути опублікована в електронному архіві відкритого доступу ElArKhNURE. Всі запозичення з друкованих та електронних джерел мають відповідні посилання.

Я ознайомлений з діючим положенням «Про протидію академічному плагіату в ХНУРЕ», згідно з яким виявлення плагіату є підставою для відмови в допуску кваліфікаційної роботи до захисту та застосування дисциплінарних заходів.

ЗМІСТ

ВСТУП

У сучасному світі соціальні мережі стали важливою частиною життя людей, виконуючи ключові комунікативні, інформаційні та навіть освітні функції. Вони сприяють не тільки розвитку міжособистісних відносин, але і дозволяють створювати цілі спільноти, об'єднані спільними інтересами, цінностями або ідеями. Розуміння структури та динаміки таких спільнот у соціальних мережах має велике значення для науковців, практиків та організацій, оскільки воно може допомогти виявити закономірності поведінки користувачів, зрозуміти вплив різних факторів на формування спільнот, а також розробити стратегії для впровадження ефективних маркетингових кампаній, комунікаційних програм та інших ініціатив.

В цьому контексті аналіз кластеризації стає потужним інструментом для дослідження спільнот у соціальних мережах. Кластеризація - це процес групування об'єктів на основі їхньої схожості чи відмінності з метою виявлення прихованих структур або закономірностей. У випадку соціальних мереж кластеризація може допомогти виявити групи користувачів зі схожими інтересами, поведінкою або взаємодіями, що може свідчити про наявність стійких спільнот.

Дослідження методів кластеризації даних при аналізі спільнот у соціальних мережах має велику актуальність та практичну значимість. У рамках цієї роботи будуть розглянуті різні методи кластеризації, їх переваги та недоліки, а також виберемо найбільш підходящі методи для аналізу спільнот у соціальних мережах. В роботі розглянуті такі алгоритмах кластеризації, як ієрархічна кластеризація, k-середніх, DBSCAN, модулярність та інші, що демонструють гарні результати у вирішенні подібних задач.

Одним з ключових аспектів дослідження є вибір відповідних даних для аналізу. Соціальні мережі надають велику кількість інформації про користувачів, їх взаємодії та поведінку, проте не всі дані можуть бути корисними для аналізу спільнот. Особливо важливим є вибір атрибутів, які дозволять визначити схожість між користувачами та виявити закономірності в їхньому об'єднанні в спільноти.

У практичній частині дослідження застосовані методи кластеризації до реальних даних з соціальної мережі Facebook. В роботі проаналізована

ефективність різних методів та порівняємо результати кластеризації, отримані за їх допомогою. Для цього були визначені метрики успішності, які дозволяють оцінити якість кластерів та виявлення спільнот. Деякі з таких метрик можуть включати забезпечення внутрішньої однорідності кластерів (схожість між елементами кластера), зовнішню роздільність (відмінність між різними кластерами) та стабільність кластерів (збереження структури кластерів після невеликих змін у даних).

Для проведення аналізу використовуються різні інструменти та бібліотеки, доступні для мов програмування Python, такі як NetworkX, scikit-learn, Gephi тощо. Ці інструменти дозволяють працювати з графами мережі, застосовувати алгоритми кластеризації та візуалізувати результати дослідження.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ

1.1 Особливості соціальних мереж

У сучасному світі соціальні мережі стали важливою частиною життя багатьох людей. Вони допомагають користувачам спілкуватися, обмінюватися інформацією, знаходити однодумців і розвивати свої професійні навички. Соціальні мережі є масштабними та складними системами, які щодня зберігають та обробляють величезний обсяг даних [1, 2]. Аналіз цих даних може допомогти вирішити ряд практичних проблем, зокрема у сфері реклами, маркетингу, рекомендаційних систем тощо.

Одним із ключових напрямків аналізу соціальних мереж є дослідження спільнот, що формуються серед їхніх користувачів. Спільноти можуть базуватися на різних критеріях, таких як спільні інтереси, географічне розташування, професійні зв'язки тощо. Виявлення спільнот та аналіз їх структури та характеристик можуть допомогти виявити ключових учасників, прослідкувати розповсюдження інформації або визначити особливості взаємодії користувачів у мережі.

Соціальні мережі використовують різні моделі даних, в залежності від своєї природи та вимог до функціональності. Основними типами моделей даних в соціальних мережах є графові моделі, реляційні моделі та моделі на основі документів.

Графова модель – це основний метод представлення даних в соціальних мережах. У графовій моделі користувачі представлені у вигляді вузлів, а відносини між ними – у вигляді ребер, що з'єднують вузли. У такому випадку соціальна мережа може бути визначена як граф G = (V, E), де V – це множина вузлів (користувачів), а E – множина ребер (відносин) [1].

Графові моделі дозволяють виконувати складні запити на знаходження шляхів, підрахунок ступеня вузлів, аналіз центральності та інші аналітичні операції

Реляційні моделі даних були розроблені для того, щоб структурувати та зберігати дані в табличному вигляді. Вони використовують сутності та відносини між ними для представлення даних. Кожна сутність відображається у вигляді

таблиці, де рядки представляють окремі записи, а стовпці представляють атрибути сутності. Відносини між сутностями встановлюються за допомогою зовнішніх ключів.

Хоча реляційні моделі даних можуть бути не такими гнучкими та ефективними для моделювання складних взаємозв'язків, як графові моделі, вони забезпечують добре структуровані та надійні механізми для зберігання та обробки даних., що важливі для аналізу соціальних мереж.

Моделі на основі документів представляють дані у вигляді документів, що включають ключові значення та гнучкі структури даних. Ці моделі, зазвичай, використовують формати, такі як JSON або XML, для зберігання даних. Вони надають більше гнучкості, ніж реляційні моделі, та можуть легко пристосовуватися до змін в структурі даних [1].

Однією з переваг моделей на основі документів є те, що вони дозволяють виконувати складні запити та агрегації без необхідності використовувати складний SQL. Проте, вони також мають свої обмеження, зокрема, вони можуть бути неефективними для виконання операцій, які вимагають зв'язків між документами.

Кластеризація даних в соціальних мережах вимагає досить специфічного підходу до моделювання даних. На відміну від багатьох інших типів аналізу даних, кластеризація вимагає врахування взаємозв'язків між користувачами, що робить графові моделі особливо привабливими для цієї задачі. Навпаки, реляційні моделі та моделі на основі документів можуть бути не такими ефективними для кластеризації, хоча вони можуть бути корисними для інших завдань аналізу даних [2].

Кластеризація даних є одним із основних методів аналізу безпосередньо спільнот у соціальних мережах. Вона полягає в розбитті множини об'єктів на групи (кластери) таким чином, щоб об'єкти всередині кластера були схожими між собою, а об'єкти з різних кластерів – відрізнялися. Застосування кластеризації до аналізу соціальних мереж допомагає виявляти спільноти, розпізнавати їх структуру та динаміку розвитку. Однак, через особливості соціальних мереж, такі як велика кількість користувачів, змінюваність структури взаємодій, гетерогенність даних та

наявність шуму, застосування традиційних методів кластеризації може бути недостатньо ефективним. Тому актуальним є дослідження нових підходів та методів кластеризації, які б були пристосовані до особливостей соціальних мереж.

Актуальність теми дослідження зумовлена науковою та практичною потребою в розробці та вдосконаленні методів кластеризації для аналізу спільнот у соціальних мережах. Соціальні мережі є потужним інструментом для вивчення соціальних процесів, розуміння тенденцій та прогнозування поведінки користувачів. Отже, розробка ефективних методів аналізу спільнот може мати значний вплив на різні сфери, зокрема бізнес, політику, науку та освіту.

З урахуванням вищевикладеного, можна зробити висновок, щіальних мережах" є актуальною та важливою для наукової спільноти і суспільства в цілому. Результати дослідження можуть сприяти розвитку нових підходів до аналізу та виявлення спільнот, що забезпечать краще розуміння структури та динаміки соціальних мереж. Це може мати позитивний вплив на ряд сфер діяльності, зокрема маркетинг, рекламу, управління спільнотами, рекомендаційні системи та наукові дослідження.

Подальше дослідження методів кластеризації та їх застосування для аналізу спільнот у соціальних мережах може допомогти виявити нові можливості та підходи для розробки ефективних алгоритмів, забезпечити більш точний та оперативний аналіз взаємодій користувачів та виявлення ключових учасників спільнот. Окрім того, результати дослідження можуть бути використані для виявлення маніпуляцій, впливу на громадську думку та інших аспектів, які стосуються безпеки та прозорості соціальних мереж.

Таким чином, дослідження методів кластеризації даних при аналізі спільнот у соціальних мережах відповідає актуальним науковим та практичним вимогам, має значний потенціал для подальшого розвитку.

1.2 Особливості спільнот в соціальних мережах

Спільноти в соціальних мережах є фундаментальним елементом, що визначає взаємодії та комунікації між користувачами. Вони представляють собою групи

користувачів, які мають спільні інтереси, цінності або відносини. Спільноти можуть бути формальними (наприклад, групи в Facebook або LinkedIn) або неформальними, що виражається в невиразних паттернах взаємодії між користувачами [2].

Соціальні мережі, особливо великі, такі як Facebook, Twitter або LinkedIn, містять велику кількість даних, які можна аналізувати для розуміння структури та динаміки спільнот. Величезний обсяг даних, які генеруються користувачами соціальних мереж щодня, вимагає використання розширюваних методів аналізу даних для виявлення цінностей. Дані, зібрані з соціальних мереж, мають велику вимірність, яка включає текст, зображення, відео, метадані та інше. На рисунку 1.1 можна побачити ескіз невеликої мережі, що відображає структуру спільноти з трьома групами вузлів із щільними внутрішніми зв’язками та рідшими зв’язками між групами.

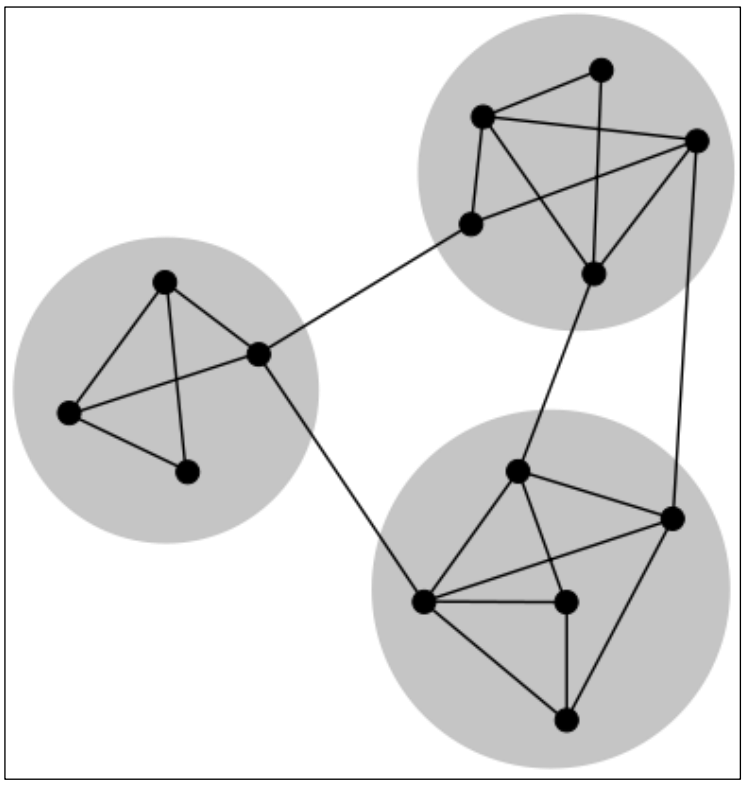

Рисунок 1.1 – Структура спільнот в соціальних мережах [2]

Важливим аспектом аналізу спільнот в соціальних мережах є виявлення спільнот, або кластеризація. Це процес групування вузлів у кластери таким чином, що вузли всередині кластера мають високу ступінь схожості, тоді як вузли з різних кластерів мають низьку ступінь схожості [2 []]. Кластеризація може бути основана

на різних критеріях, таких як спільні інтереси, відносини, демографічні характеристики тощо.

Кластеризація в соціальних мережах має великий потенціал для розуміння і виявлення складних взаємозв'язків між користувачами та виявлення спільнот зі схожими інтересами або характеристиками. Вона може допомогти у виявленні груп людей, які мають схожі думки, поведінку або цілі, що може мати важливе значення для багатьох сфер життя, таких як маркетинг, рекомендаційні системи, соціальна аналітика та ін.

Однак кластеризація даних з соціальних мереж може бути викликом через великий обсяг та вимірність даних, а також через динаміку мережі. Тому використання ефективних алгоритмів кластеризації, які можуть швидко та точно обробляти великі набори даних, є важливим.

Науковий аналіз даних з соціальних мереж може допомогти у виявленні важливих моделей взаємодії, прогнозуванні майбутніх тенденцій, виявленні впливових користувачів, визначенні спільнот інтересів та багато іншого. Завдяки впровадженню методів машинного навчання та штучного інтелекту можливо виробити більш глибоке та інсайтове розуміння структур та процесів у соціальних мережах.

Для досягнення кращих результатів у кластеризації соціальних мереж, можна поєднувати методи машинного навчання з графовими алгоритмами, що дозволяє ефективно моделювати та прогнозувати поведінку користувачів у мережі. Комбінація цих підходів може привести до знаходження нових підходів та узагальнених моделей, які допоможуть зрозуміти складні динамічні процеси, що відбуваються у соціальних мережах [3].

Таблиця 1.1 – Ключові характеристики спільнот в соціальних мережах [3]

Характеристика	Опис

Масштаб	Соціальні мережі можуть містити мільйони або навіть мільярди користувачів, що створює виклики для аналізу даних.
Висока вимірність	Дані з соціальних мереж можуть включати текст, зображення, відео, метадані, що вимагає використання високовимірних методів аналізу.
Динаміка	Соціальні мережі є динамічними, що вимагає використання методів, які можуть враховувати зміни в структурі та взаємодії мережі.
Геометрія даних	Структура спільноти може бути представлена у вигляді графів, які мають свої особливості і виклики для аналізу.
Шум і неповнота	Дані з соціальних мереж часто містять шум і можуть бути неповними, що вимагає використання робустих методів обробки даних.

Ще однією особливістю соціальних мереж є те, що вони не просто збірник даних, а система, в якій інформація та вплив поширюються через мережу взаємодій. Така динаміка може бути важливим аспектом при аналізі спільнот. Наприклад, важливим завданням може бути ідентифікація впливових членів спільноти, які можуть вплинути на розповсюдження інформації або поведінку інших членів спільноти.

Виклики в аналізі спільнот в соціальних мережах, такі як великий обсяг даних, висока вимірність і динаміка, вимагають використання продвинутих методів аналізу даних та машинного навчання. Однак, враховуючи ці виклики, можна видобути цінні інсайти з даних, що можуть бути використані для різних цілей, від наукового дослідження до маркетингу.

У додаток до викликів, аналіз спільнот в соціальних мережах також має потенціал для дослідження нових явищ і динаміки. Наприклад, можна досліджувати, як формуються та розвиваються спільноти, як інформація та вплив

розповсюджуються через мережу, які фактори впливають на активність користувачів, і так далі.

1.3 Огляд кластеризації даних

Кластеризація є одним із важливих методів дослідження в машинному навчанні та аналізі даних, який використовується для розподілу об'єктів на групи або кластери на основі подібності. Мета кластеризації полягає у тому, щоб об'єкти в одному кластері були схожі між собою і відрізнялися від об'єктів в інших кластерах.

Кластеризація даних використовується в різних сферах і допомагає вирішувати низку задач. Наприклад, в соціальних мережах кластеризація може допомогти ідентифікувати групи користувачів з подібними інтересами, виявити незвичайну поведінку або виявити спільноти. В маркетингу та рекламі кластеризація може бути використана для сегментації ринку, визначення цільових груп або прогнозування поведінки споживачів [3].

Кластеризація може вирішити багато проблем аналізу даних, особливо коли стикаємося з великими обсягами даних, такими як соціальні мережі. Часто, коли дані надто об'ємні, традиційний аналіз може бути вкрай складним або неможливим. Тут кластеризація може надати можливість розбити даних на групи, що дозволяє більш зосередитись на конкретних об'єктах аналізу.

З іншого боку, кластеризація також може допомогти знайти скриті структури або шаблони в даних. Це особливо корисно в соціальних мережах, де можна виявити спільноти або групи користувачів з подібними інтересами або поведінкою.

Однак, кластеризація даних не є простим завданням і має ряд викликів. Наприклад, вибір відповідного алгоритму кластеризації і кількості кластерів може бути складним, особливо для неструктурованих або великих наборів даних.

Крім того, результати кластеризації можуть залежати від вибраної міри схожості або відстані, що може впливати на якість кластерів. Наприклад, використання евклідової відстані може бути недоцільним для даних високої розмірності через так зване «прокляття розмірності».

Також існує проблема інтерпретації результатів кластеризації. Оскільки кластеризація є методом неконтрольованого навчання, немає «правильних відповідей» для порівняння з отриманими кластерами. Оцінка якості кластеризації і визначення, чи відображають кластери реальні закономірності в даних, можуть бути складними завданнями. Нижче розглянуті поширені області застосування, в яких виникає проблема кластеризації [4].

Спільна фільтрація – у методах спільної фільтрації кластеризація забезпечує узагальнення користувачів-однодумців. Оцінки, надані різними користувачами один одному, використовуються для виконання спільної фільтрації. Це можна використовувати для надання рекомендацій у різноманітних програмах.

Сегментація клієнтів – ця програма дуже схожа на спільне фільтрування, оскільки створює групи подібних клієнтів у даних. Основна відмінність від спільної фільтрації полягає в тому, що замість використання рейтингової інформації для цілей кластеризації можуть використовуватися довільні атрибути об’єктів.

Узагальнення даних – багато методів кластеризації тісно пов’язані з методами зменшення розмірності. Такі методи можна вважати формою узагальнення даних. Узагальнення даних може бути корисним у створенні компактних представлень даних, які легше обробляти та інтерпретувати в різноманітних програмах.

Динамічне виявлення трендів – багато форм динамічних і потокових алгоритмів можна використовувати для виявлення трендів у різноманітних програмах соціальних мереж. У таких програмах дані динамічно кластеризуються потоковим способом і можуть використовуватися для визначення важливих моделей змін. Прикладами таких потокових даних можуть бути багатовимірні дані, текстові потоки, потокові дані часових рядів і дані траєкторії. Ключові тенденції та події в даних можна виявити за допомогою методів кластеризації.

Аналіз мультимедійних даних – низка різних видів документів, таких як зображення, аудіо чи відео, підпадають під загальну категорію мультимедійних даних. Визначення подібних сегментів має численні застосування, наприклад,

визначення подібних фрагментів музики чи подібних фотографій. У багатьох випадках дані можуть бути мультимодальними та містити різні типи. У таких випадках проблема стає ще складнішою.

А також аналіз соціальної мереж – у цих програмах структура соціальної мережі використовується для визначення важливих спільнот у базовій мережі. Виявлення спільноти має важливе застосування в аналізі соціальних мереж, оскільки воно забезпечує важливе розуміння структури спільноти в мережі. Кластеризація також має застосування для підсумовування соціальних мереж, що корисно в ряді програм.

В контексті цього дослідження, кластеризація буде використовуватись для виявлення спільнот в соціальних мережах. Це дозволить краще зрозуміти структуру та динаміку соціальних мереж, а також ідентифікувати ключові характеристики і патерни поведінки користувачів. Далі, будуть оглянуті декілька методів кластеризації і порівняння їх ефективність.

1.4 Постановка задач та основні завдання дослідження

Метою даного дослідження є розробка та вдосконалення методів кластеризації даних для аналізу спільнот у соціальних мережах, які враховують особливості цих мереж, забезпечують високу точність та ефективність аналізу, а також масштабуються для обробки великих обсягів даних. Для досягнення цієї мети, необхідно вирішити ряд конкретних завдань:

- вивчення теоретичних основ кластеризації даних, включаючи поняття, методи та алгоритми кластеризації, які застосовуються для аналізу спільнот у соціальних мережах;
- огляд сучасних алгоритмів кластеризації, які застосовуються для аналізу спільнот у соціальних мережах, та визначення їх переваг та недоліків;
- вивчення особливостей застосування кластеризації в аналізі спільнот у соціальних мережах, а також основних викликів та проблем, пов'язаних з цим процесом;

- розробка вдосконалених методів кластеризації, які враховують специфіку соціальних мереж, можуть масштабуватися для обробки великих даних та забезпечують високу точність та ефективність аналізу спільнот;
- проведення експериментів з використанням вдосконалених методів кластеризації на даних соціальних мереж, оцінка результатів та порівняння їх з існуючими методами.
- формулювання висновків щодо ефективності вдосконалених методів кластеризації та їх перспективності для подальшого застосування та розвитку в аналізі спільнот у соціальних мережах.
- розробка рекомендацій щодо практичного застосування розроблених або вдосконалених методів кластеризації для аналізу спільнот у соціальних мережах.

Виконання цих завдань дозволить розв'язати актуальну проблему розробки та вдосконалення методів кластеризації даних для аналізу спільнот у соціальних мережах, що враховують особливості цих мереж, забезпечують високу точність та ефективність аналізу, а також масштабуються для обробки великих обсягів даних. У результаті дослідження можуть бути отримані нові знання та підходи, що сприятимуть підвищенню якості аналізу спільнот у соціальних мережах та розширенню можливостей використання даних мереж для наукових, практичних та комерційних цілей.

1.5 Об'єкт та предмет дослідження

Об'єкт дослідження – це процеси аналізу спільнот у соціальних мережах. Об'єкт охоплює різні аспекти взаємодії користувачів соціальних мереж, структуру спільнот, характеристики поведінки користувачів, а також методи та інструменти, що використовуються для збору, обробки та аналізу даних про спільноти.

Предмет дослідження – методи кластеризації даних, які застосовуються для аналізу спільнот у соціальних мережах. В рамках предмета дослідження вивчаються теоретичні основи кластеризації, сучасні алгоритми кластеризації, особливості застосування кластеризації в аналізі спільнот у соціальних мережах, а

також розробка та вдосконалення нових методів кластеризації, які враховують специфіку соціальних мереж, масштабуються для обробки великих даних та забезпечують високу точність та ефективність аналізу.

Виконання завдань, визначених в підрозділі 1.2, спрямоване на розв'язання актуальних проблем, пов'язаних з аналізом спільнот у соціальних мережах, а також на отримання нових знань та підходів, що сприятимуть підвищенню якості аналізу спільнот та розширенню можливостей використання даних мереж для наукових, практичних та комерційних цілей.

2 АНАЛІЗ СПІЛЬНОТ У СОЦІАЛЬНИХ МЕРЕЖАХ

2.1 Задачі аналізу соціальних мереж

Соціальні мережі – це величезні сховища інформації про поведінку користувачів, структуру взаємодії та інші аспекти соціальних відносин. Аналіз соціальних мереж відкриває нові можливості для вивчення динаміки та закономірностей соціальних систем [4, 5]. Кластеризація допомагає виявляти закономірності та структури в цих масивах даних, що може сприяти розвитку нових стратегій, продуктів або підходів до взаємодії з аудиторією.

Аналіз соціальних мереж є складною та розмаїтою задачею, оскільки він представляє собою великі та складні системи з великою кількістю взаємодіючих об'єктів, таких як користувачі, групи, спільноти та взаємозв'язки між ними. Аналіз соціальних мереж став популярним напрямком досліджень через його потенціал у вивченні соціальних відносин, виявленні спільнот та розумінні динаміки цих мереж.

Одним з основних завдань аналізу соціальних мереж є виявлення спільнот користувачів. Спільнота – це група користувачів, які мають більш тісні зв'язки між собою, ніж з рештою мережі. До цієї категорії відносяться, наприклад, друзі, колеги, члени сім'ї, люди з однаковими інтересами та ін. Задача ідентифікації спільнот має важливе значення в контексті реклами, рекомендаційних систем, а також аналізу поведінки користувачів.

Соціальні мережі містять велику кількість відносин між користувачами, які відображають різні аспекти їх взаємодії. Аналіз зв'язків допомагає виявити ключових учасників, визначити структуру мережі, знайти центри впливу та виявити зміни в структурі мережі з часом [5].

Важливими вузлами є користувачі, які мають високу центральність, велику кількість зв'язків або відіграють ключову роль в процесах поширення інформації в мережі. Виявлення таких вузлів може бути корисним для рекламних кампаній, виявлення лідерів думок та аналізу процесів поширення інформації.

Соціальні мережі – це основний канал поширення новин, ідей, мемів та інших видів інформації. Аналіз поширення інформації допомагає зрозуміти, як

інформація поширюється в мережі, хто відіграє ключову роль в цьому процесі, та які фактори впливають на швидкість і обсяг поширення інформації.

За допомогою аналізу соціальних мереж можна виявити закономірності в поведінці користувачів і використовувати ці знання для прогнозування їх майбутньої поведінки. Це може бути особливо корисно для рекомендаційних систем, рекламних кампаній та інших застосувань.

Завдання кластеризації полягає в розділенні користувачів на групи за різними критеріями. За допомогою кластерного аналізу можна виявити групи користувачів зі схожими інтересами, поведінкою або іншими ознаками.

Всі ці задачі вимагають використання різних методів аналізу даних, включаючи методи машинного навчання, статистичні методи, методи графової теорії та ін. Вони також вимагають вміння працювати з великими обсягами даних, обробляти шумові дані, а також враховувати динаміку мережі та поведінку користувачів.

2.2 Збір даних

Збір даних – це критично важливий етап для будь-якого дослідницького проекту. У контексті аналізу соціальних мереж цей процес стає особливо складним і непередбачуваним через величезний обсяг даних, що генеруються користувачами, та динаміку їх змін. До того ж, кожна соціальна мережа має свої власні особливості, що вимагає від дослідника глибокого розуміння технологій та протоколів збору даних. Далі буде розгляншутий збір даних з різних соціальних мереж.

Facebook Graph API [6] – це потужний інструмент, що дозволяє дослідникам отримувати доступ до великої кількості даних. Ви можете отримувати дані про пости, коментарі, реакції, фотографії, відео, учасників груп та багато іншого. Але є важливе обмеження: вам потрібно отримати дозвіл на доступ до даних. В залежності від типу даних, до яких ви намагаєтеся отримати доступ, ви повинні створити програму та отримати від Facebook спеціальний токен доступу. Незважаючи на це, Facebook Graph API – це чудовий інструмент для збору даних з однієї з найбільших соціальних мереж (рис.2.1).

Рисунок 2.1 – Facebook Graph API

Twitter надає дослідникам доступ до Twitter API [7], що дозволяє збирати різні типи даних. Це включає публічні твіти, інформацію про користувача, інформацію про фолловерів та following, а також детальну інформацію про окремі твіти, включаючи ретвіти, відповіді, лайки та інше. Для отримання доступу до Twitter API вам потрібно створити програму на платформі розробника Twitter, а потім отримати ключі та токени для авторизації вашої програми. Twitter API дозволяє збирати великі обсяги даних в режимі реального часу, що робить його чудовим інструментом для збору даних для аналізу соціальних мереж.

Instagram API [8], в порівнянні з Facebook Graph API та Twitter API, має деякі обмеження. Instagram Basic Display API дозволяє дослідникам отримувати основну інформацію про користувача, його пости та фолловерів, але не дозволяє збирати коментарі, лайки або відповіді. Для отримання доступу до Instagram API, як і в інших випадках, вам потрібно створити програму та отримати токен доступу (рис.2.2).

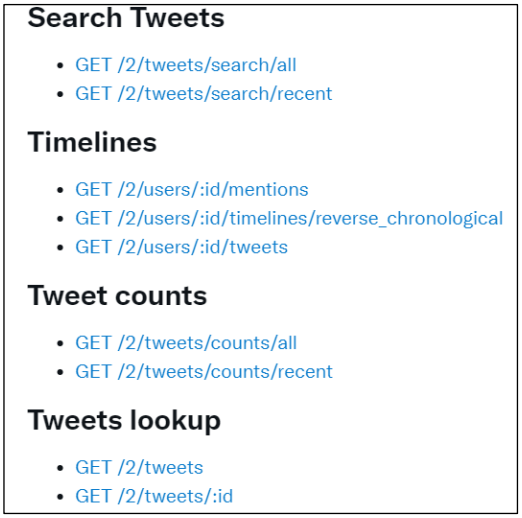

Рисунок 2.2 – Приклад запитів Twitter API

LinkedIn API [9] надає можливість збирати інформацію про профілі користувачів, їх зв'язки, досвід роботи, освіту, навички, рекомендації та багато іншого. Також доступні дані про компанії, вакансії та групи. Для доступу до LinkedIn API вам потрібно створити програму на платформі розробників LinkedIn і отримати токен доступу.

Reddit також має свій API [10], який дозволяє збирати дані про пости, коментарі, голосування, інформацію про користувачів та багато іншого. Reddit API є дуже гнучким і дозволяє вам збирати дані з конкретних subreddit-ів, що робить його ідеальним для збору цільових даних для аналізу спільнот. Для доступу до Reddit API вам потрібно створити програму на платформі розробників Reddit та отримати токен доступу.

Збір даних з цих соціальних мереж є тільки першим кроком. Після того, як дані зібрані, вони повинні бути оброблені та аналізовані, що часто є значно складнішою задачею.

2.2 Розв'язання задач аналізу

В контексті соціальних мереж, кластеризація часто використовується для виявлення спільнот, які мають схожі інтереси або погляди. Це може бути корисно для багатьох цілей, включаючи таргетовану рекламу, розробку продуктів, рекомендаційні системи та інше.

Кластеризація в соціальних мережах здійснюється шляхом аналізу різних типів взаємодій користувачів. Це можуть бути взаємодії користувач-користувач, такі як дружба або спільні коментарі, або взаємодії користувач-контент, такі як лайки або репости. Такі взаємодії дозволяють визначити схожість між користувачами, що є основою для кластеризації. ля виявлення спільнот у соціальних мережах можна використовувати методи кластеризації [5, 11].

Один з найпоширеніших методів – K-means, розділяє користувачів на групи, засновані на подібності їхніх характеристик або взаємодій. Додатково, алгоритми на основі щільності, такі як DBSCAN, виявляють густі регіони у графі соціальних зв'язків. Ієрархічні методи, такі як Agglomerative Clustering, будують дерево об'єднань, що дозволяє виділити спільноти на різних рівнях агрегації. Методи змішування, такі як Gaussian Mixture Models, моделюють розподіл даних та виявляють спільноти, які відповідають різним компонентам розподілу.

Для аналізу впливу у соціальних мережах також можна використовувати методи кластеризації. Кластери, які містять користувачів з високим рівнем впливу, можуть бути виявлені за допомогою алгоритмів кластеризації, які групують користувачів за характеристиками, що свідчать про їхній вплив, такі як кількість друзів, підписників або частота публікацій. Кластери також можуть відображати взаємодію між впливовими користувачами та їхніми прихильниками, допомагаючи розкрити динаміку впливу у мережі.

Для аналізу емоцій та настроїв у соціальних мережах методи кластеризації можуть бути використані для групування користувачів або текстових даних за емоційними характеристиками. Застосування методів, таких як K-means, DBSCAN або агломеративна кластеризація, дозволяє виявити групи користувачів з схожими емоційними реакціями на певні події або контент. Це дає можливість організаціям аналізувати емоційну реакцію своєї аудиторії, розуміти їхні настрої та приймати відповідні рішення щодо комунікації та маркетингу.

Кластеризація використовується в рекомендаційних системах для групування користувачів або об'єктів на основі їхніх інтересів, поведінки та характеристик. Застосування методів, таких як K-means, DBSCAN або графові

методи кластеризації, дозволяє розробляти рекомендаційні системи, які пропонують користувачам відповідний контент або продукти, засновуючись на їхніх інтересах та схожості з іншими користувачами. Це дозволяє покращити персоналізацію рекомендацій та задовольнити потреби користувачів.

Кластеризація також може бути використана для сегментації аудиторії на основі різних ознак, таких як вік, стать, географічне положення, інтереси тощо. Застосування методів кластеризації, таких як K-means, DBSCAN або ієрархічна кластеризація, дозволяє визначити різні сегменти аудиторії, що допомагає маркетологам краще розуміти своїх клієнтів та налаштовувати маркетингові стратегії для кожної групи. Це дозволяє створювати більш цільові та ефективні рекламні кампанії, що забезпечують більш високу конверсію та задоволення клієнтів.

Ієрархічна кластеризація – це ще один підхід, який може бути використаний для аналізу соціальних мереж. Ієрархічна кластеризація використовує дендрограми для представлення структури спільноти та допомагає визначати рівень деталізації, який необхідний для кожної конкретної задачі аналізу. Існує два основних види ієрархічної кластеризації: агломеративна (об'єднання) та дивізивна (розщеплення). Агломеративна кластеризація починає з окремих вершин графа та поступово об'єднує їх у кластери, засновуючись на мірах відстані між ними. Натомість дивізивна кластеризація розпочинається з одного великого кластера, який містить усі вершини, та рекурсивно розділяє його на менші кластери.

Для розв'язання задач аналізу соціальних мереж використовуються інструменти та платформи, які спрощують процес дослідження та візуалізацію даних. Деякі популярні інструменти та бібліотеки:

- Gephi – це відкритий інструмент для візуалізації та аналізу графів, який допомагає дослідникам виявляти закономірності, структуру та взаємодію у графах [11]. Він надає різні алгоритми виявлення спільнот, міри центральності та можливість створення інтерактивних візуалізацій. Grephi дозволяє аналізувати мережеві структури, спільноти та ключових учасників, а також проектувати мережеві графіки на статичних і

динамічних мережах, геолокованих даних, а також багаторежимних/мультиплексних мережах;

- NetworkX – це потужна бібліотека для роботи з графами та мережами в мові програмування Python [11]. Вона надає широкий спектр функцій для створення, модифікації та аналізу графів, а також включає різноманітні алгоритми для розв'язання завдань, пов'язаних з графовою аналітикою. NetworkX також пропонує реалізацію різних алгоритмів, таких як пошук найкоротших шляхів, алгоритми кластеризації, алгоритми пошуку підграфів та інші. Це дозволяє виконувати складні завдання аналізу мереж, включаючи виявлення спільнот, впливовий аналіз, виявлення центральних вузлів та багато іншого. Бібліотека NetworkX також має простий та зрозумілий інтерфейс, що робить її легкою у використанні та розширенні. Вона інтегрується з іншими популярними бібліотеками для аналізу даних в Python, такими як NumPy та Pandas, що дозволяє зручно використовувати їх разом для розв'язання складних задач аналізу соціальних мереж;
- Graph-tool – це потужна бібліотека для аналізу графів та соціальних мереж [11]. Вона надає широкий набір функцій для роботи з графами, включаючи конструювання, модифікацію, візуалізацію та аналіз графових структур. Однією з головних переваг Graph-tool є його швидкодія. Вона використовує ефективні алгоритми та оптимізовані структури даних, що дозволяє обробляти навіть дуже великі графи з мільйонами вершин та ребер. Це особливо важливо для аналізу соціальних мереж, які можуть мати складну структуру та великі розміри даних. Graph-tool підтримує широкий спектр алгоритмів для аналізу графів, включаючи кластеризацію, пошук шляхів, обчислення центральності та багато інших. Крім того, бібліотека має зручний та інтуїтивно зрозумілий інтерфейс програмування, що спрощує використання та розробку аналітичних рішень;

- scikit-learn – це відкрита бібліотека машинного навчання для мови програмування Python [11]. Вона надає широкий спектр алгоритмів інтелектуального аналізу даних, включаючи класифікацію, регресію, кластеризацію, зменшення розмірності, виявлення аномалій та багато іншого. Sklearn реалізує багато популярних алгоритмів машинного навчання, таких як метод опорних векторів (SVM), рішучі дерева, нейронні мережі, алгоритми кластеризації, наприклад, K-means та DBSCAN, та багато інших. Бібліотека також надає інструменти для підготовки даних, оцінки моделей, використання перехресної перевірки та налаштування параметрів моделей. Однією з переваг Sklearn є його простота використання та документація. Вона надає зрозумілі та детальні приклади використання кожного алгоритму, що допомагає швидко освоїти інструменти машинного навчання;
- NLTK (Natural Language Toolkit) та spaCy – це бібліотеки Python для обробки природної мови, які допомагають виявляти ключові слова, теми та емоції, аналізувати текстові дані та будувати моделі машинного навчання для розпізнавання мовних закономірностей [11]. NLTK надає зручний інтерфейс та широкий спектр функцій для роботи з текстовими даними, такими як токенізація, стемінг, лематизація, визначення частин мови, аналіз синтаксису та багато іншого. NLTK також включає в себе колекцію корпусів текстових даних, лексиконів та моделей машинного навчання для різних завдань обробки природньої мови. SpaCy – це швидка та ефективна бібліотека для обробки природньої мови (NLP) також для мови програмування Python. Вона спеціалізується на виконанні швидкого та точного аналізу тексту, включаючи токенізацію, лематизацію, визначення частин мови, залежностей та іменованих сутностей.

Ці інструменти та бібліотеки можуть бути комбіновані та адаптовані для різних задач аналізу соціальних мереж, від простого візуального представлення графів до складного аналізу даних та створення передбачувальних моделей. Вибір

підходу, методу аналізу та інструментів залежить від конкретних потреб дослідника та особливостей досліджуваної соціальної мережі.

2.3 Проблеми існуючих рішень

У розділі був оглянутий ряд існуючих методів та інструментів для аналізу спільнот у соціальних мережах. Однак, ці рішення мають свої обмеження та недоліки, які визначають потребу в дослідженні ефективності застосування методів кластеризації у цьому контексті.

Багато існуючих методів аналізу спільнот у соціальних мережах засновані на графових алгоритмах та кластеризації. Однак, кластеризація на основі графів може мати низку обмежень [12]:

- вибір оптимальної кількості кластерів. Визначення оптимальної кількості кластерів може бути непростою задачею, оскільки вона залежить від специфіки даних та мети дослідження. Неконтрольовані методи кластеризації, такі як K-means, потребують заздалегідь заданої кількості кластерів, що може призводити до підбору параметрів та неоптимальних результатів.
- чутливість до шуму та викидів. Графові методи кластеризації можуть бути чутливими до шуму та викидів у даних, що може призводити до неправильного виявлення спільнот або впливати на якість кластеризації. Деякі методи, такі як DBSCAN, можуть враховувати шум, але все ще можуть мати складнощі з виявленням слабо структурованих спільнот.
- часова та обчислювальна складність. Великі графи та мережі можуть мати мільйони вершин та ребер, що може призвести до високої часової та обчислювальної складності деяких графових алгоритмів кластеризації. Це може обмежувати їх застосування у великомасштабних дослідженнях та реальних сценаріях.
- нестабільність результатів. Результати кластеризації можуть бути нестабільними або залежними від початкових умов або параметрів алгоритму. Це може призводити до різних розбиттів графа та виявлення

спільнот в залежності від випадкових факторів, що може ускладнювати порівняння та аналіз результатів.

Існуючі методи аналізу спільнот у соціальних мережах можуть мати обмеження, які залежать від специфіки мережі, типу даних або мети дослідження. Наприклад, методи кластеризації, засновані на структурі графа, можуть бути неефективними для аналізу спільнот, які засновані на атрибутах користувачів або їхніх поведінкових особливостях. Це підкреслює потребу в розвитку універсальних методів, які можуть адаптуватися до різних ситуацій та вимог.

Оцінка якості кластеризації та виявлення спільнот може бути складною задачею через відсутність універсальних метрик та золотих стандартів. Деякі метрики, такі як модулярність або нормалізована мутуальна інформація, можуть бути використані для оцінки результатів, але вони можуть мати свої недоліки та обмеження. Наприклад, модулярність може мати тенденцію до переваги великих кластерів або сприяти розділенню спільнот, які перетинаються. Також, часто відсутній «золотий стандарт» для порівняння результатів різних методів, що може ускладнювати вибір найкращого підходу та аналіз результатів.

Соціальні мережі є динамічними структурами, які постійно змінюються через появу нових користувачів, взаємодій та контенту. Багато існуючих методів аналізу спільнот можуть мати обмеження при роботі з динамічними мережами, оскільки вони передбачають статичні структури або потребують значних обчислень для адаптації до змін.

Аналіз спільнот у соціальних мережах може вимагати доступу до великої кількості персональних даних користувачів, що може порушувати приватність та конфіденційність. Це ставить питання про збалансування між ефективністю аналізу та захистом даних користувачів. Дослідження методів кластеризації, які можуть працювати з анонімізованими даними або використовувати техніки збереження приватності, є актуальним для розвитку безпечних і ефективних рішень.

3 АНАЛІЗ ІСНУЮЧИХ МЕТОДІВ КЛАСТЕРИЗАЦІЇ

3.1 Огляд понять та методів кластеризації даних

Кластеризація даних – це процес групування об'єктів з високим рівнем подібності в одній категорії або кластері, з метою зробити аналіз даних більш легким та зрозумілим. Зазвичай кластеризація використовується в таких областях, як машинне навчання, статистика, біологія та соціологія, де дані потрібно розділити на групи або підгрупи [12].

Кластер – це група об'єктів, які мають подібні властивості або характеристики. Об'єкти в кожному кластері мають схожі значення для певних параметрів, що робить їх відмінними від об'єктів інших кластерів. Кількість кластерів може бути визначена заздалегідь або знайдена за допомогою методів кластеризації.

Метод кластеризації – це підхід, який використовується для групування даних в кластери на основі певних критеріїв, які визначають подібність між об'єктами. Існує багато різних методів кластеризації, які можуть бути застосовані залежно від типу даних та вимог дослідника [12].

Два найбільш широко вивчені алгоритми кластеризації – це розділова та ієрархічна кластеризація. Ці алгоритми активно використовуються в широкому діапазоні застосувань насамперед через їхню простоту та легкість реалізації порівняно з іншими алгоритмами кластеризації. Алгоритми роздільної кластеризації спрямовані на виявлення групувань, присутніх у даних, шляхом оптимізації конкретної цільової функції та ітеративного покращення якості розділів [13]. Ці алгоритми зазвичай вимагають певних параметрів користувача для вибору точок прототипу, які представляють кожен кластер. З цієї причини їх також називають алгоритмами кластеризації на основі прототипу. Алгоритми ієрархічної кластеризації підходять до проблеми кластеризації, розробляючи структуру даних на основі бінарного дерева, яка називається дендрограмою. Після створення дендрограми можна автоматично вибрати потрібну кількість кластерів, розділивши дерево на різних рівнях, щоб отримати різні рішення кластеризації для того самого набору даних без повторного запуску алгоритму кластеризації.

Алгоритми ієрархічної кластеризації були розроблені, щоб подолати деякі недоліки, пов’язані з методами плоскої або часткової кластеризації [14]. Методи розділення зазвичай вимагають попередньо визначеного користувачем параметра K для отримання рішення кластеризації, і вони часто недетерміновані за своєю природою. Ієрархічні алгоритми були розроблені для створення більш детермінованого та гнучкого механізму для кластеризації об’єктів даних.

Ієрархічні методи можна класифікувати на агломераційні та розділові кластеризаційні методи. Агломераційні методи починаються з того, що беруть одиночні кластери (які містять лише один об’єкт даних на кластер) на нижньому рівні та продовжують злиття двох кластерів одночасно, щоб побудувати ієрархію кластерів знизу вгору. Методи розділення, з іншого боку, починаються з усіх об’єктів даних у величезному макрокластері та безперервно розділяють його на дві групи, створюючи ієрархію кластерів зверху вниз. Ієрархію кластерів тут можна інтерпретувати за допомогою стандартної термінології бінарного дерева наступним чином. Корінь представляє всі набори об’єктів даних, які потрібно кластеризувати, і це утворює вершину ієрархії (рівень 0). На кожному рівні дочірні записи (або вузли), які є підмножинами всього набору даних, відповідають кластерам [15, 16]. Записи в кожному з цих кластерів можна визначити, переходячи по дереву від поточного вузла кластера до базових однотонних точок даних. Кожен рівень в ієрархії відповідає певному набору кластерів. Основа ієрархії складається з усіх одиничних точок, які є листками дерева. Ця ієрархія кластерів також називається дендрограмою. Основна перевага ієрархічного методу кластеризації полягає в тому, що він дозволяє розрізати ієрархію на будь-якому заданому рівні та відповідно отримувати кластери. Ця функція суттєво відрізняє його від методів розділеної кластеризації тим, що він не вимагає попередньо визначеного користувачем параметра k (кількість кластерів).

Кластеризацію на основі щільності можна розглядати як непараметричний метод, оскільки він не робить припущень щодо кількості кластерів або їхнього розподілу. Кластери на основі щільності – це з’єднані щільні області в просторі даних, відокремлені одна від одної більш розрідженими областями. Крім того,

передбачається, що щільність у зонах шуму нижча, ніж щільність у будь-якому з кластерів. Завдяки своїй локальній природі щільні зв'язані області в просторі даних можуть мати довільну форму. Враховуючи структуру індексу, яка підтримує запити регіонів, кластери на основі щільності можуть бути ефективно обчислені, виконавши не більше одного запиту регіону на об’єкт бази даних. Розріджені області в просторі даних розглядаються як шум і не призначаються жодному кластеру.

3.2.1 Алгоритм K-Means

Алгоритм K-середніх є ітеративним алгоритмом кластеризації, що має на меті розбити набір даних на попередньо визначену кількість непересікаючихся підгруп (кластерів), до яких належить кожна точка даних [17, 18]. Його основна мета полягає в тому, щоб зробити точки даних в межах кластера якомога більш схожими, водночас утримуючи кластери різними (далекими) один від одного. Алгоритм починається з вибору K типових точок як початкових центроїдів. Кожна точка потім призначається найближчому центроїду на основі певної вибраної міри близькості. Після формування кластерів центроїди для кожного кластера оновлюються. Потім алгоритм ітеративно повторює ці два кроки, доки центроїди не зміняться або не буде виконано будь-який інший альтернативний критерій розслабленої конвергенції.

Алгоритм K-середніх працює за наступними кроками:

а) задати кількість кластерів K;

б) ініціалізувати центроїди, спочатку перетасовуючи набір даних і випадковим чином вибираючи точки даних для центроїдів без заміни;

в) продовжувати ітерації до тих пір, поки центроїди не перестануть змінюватися, тобто призначення точок даних до кластерів не змінюється:

1) обчислити суму квадратів відстаней між точками даних і всіма центроїдами;

2) призначити кожну точку даних до найближчого кластера (центроїда);

3) обчислити центроїди для кластерів, взявши середнє арифметичне усіх точок даних, що належать до кожного кластера.

На рисунку 3.1 показано різні етапи виконання алгоритму **3**-середніх [17]. Перша ітерація ініціалізує три випадкові точки як центроїди. У наступних ітераціях центроїди змінюють положення до збіжності. Під час обчислення найближчого центроїда в алгоритмі K-середніх можна використовувати широкий діапазон вимірювань близькості.

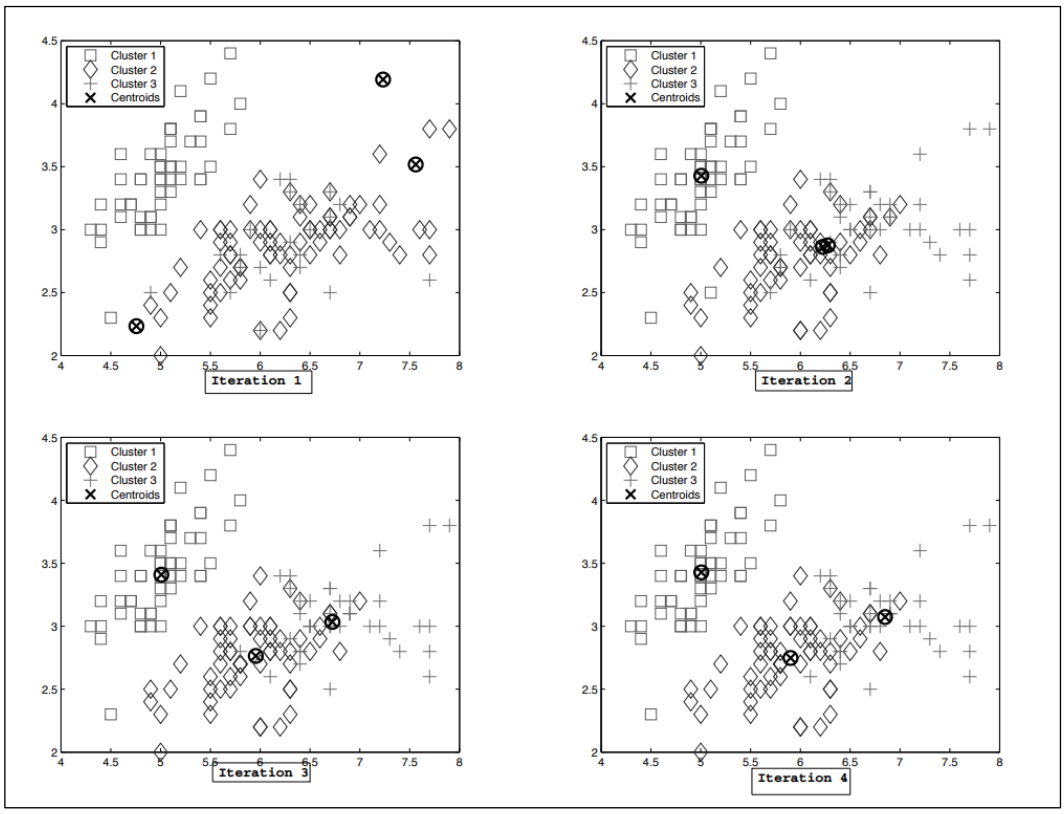

Рисунок 3.1 – 4 ітерації K-середніх [17]

Вибір може суттєво вплинути на призначення центроїда та якість остаточного рішення. Тут можна використовувати різні типи вимірювань: манхеттенська відстань (норма L_1), евклідова відстань (норма L_2) і косинус-подібність. Загалом, для кластеризації K-Means найпопулярнішим вибором є метрика евклідової відстані. Як згадувалося вище, можна отримати різні кластеризації для різних значень *K* і мір наближення. Цільова функція, яка використовується K-середніми, називається сумою квадратів помилок (SSE) або залишковою сумою квадратів (RSS).

Математичне формулювання для SSE/RSS наведено нижче [17]. Якщо набір даних $D=\{x_1,x_2,...,x_N\}$ складається з N точок, позначимо кластеризацію, отриману після застосування кластеризації K-середніх, як $C = \{C_1,C_2 ,...,C_k ...,CK\}$. SSE для цієї кластеризації визначається у рівнянні (3.1), де ck є центроїдом кластера Ck. Мета полягає в тому, щоб знайти кластеризацію, яка мінімізує оцінку SSE. Ітераційне призначення та кроки оновлення алгоритму K-Means мають на меті мінімізувати оцінку SSE для даного набору центроїдів (формули 3.1 та 3.2).

$$SSE = \sum_{k=1}^{K} \sum_{x_i \in C_k} \left|\left|x_i - c_k\right|\right|^2 (4.1) \quad (3.1)$$

$$C_k = \frac{\sum_{x_i \in C_k} x_i}{|C_k|} \quad (3.2)$$

Переваги та недоліки алгоритму K-середніх розглянуті у таблиці 3.2.

Таблиця 3.2 – Переваги та недоліки алгоритму K-середніх [18]

Переваги	Недоліки
Простота реалізації	Необхідність вибору кількості кластерів K
Швидкість та ефективність	Чутливість до початкової ініціалізації центроїдів
Можливість працювати з великими наборами даних	Відсутність гарантії знаходження глобального мінімума функції втрат
Працює добре для кластерів, що мають кулясту форму	Складнощі з кластерами

Алгоритм K-середніх є широко використовуваним методом кластеризації, що пропонує просте та ефективне рішення для різних завдань. Проте, слід звернути увагу на його обмеження та обрати відповідні параметри для конкретної задачі.

3.2.2 Алгоритм DBSCAN

DBSCAN оцінює щільність шляхом підрахунку кількості точок в околиці з фіксованим радіусом і вважає дві точки зв’язаними, якщо вони знаходяться в околицях одна одної. Точка називається основною, якщо околиця радіуса *Eps* містить принаймні точки *MinPts*, тобто щільність у околиці має перевищувати деякий поріг [17, 18]. Точка q є прямою щільністю досяжною з основної точки p, якщо q знаходиться в *Eps*-околиці p, а щільність досяжності задана транзитивним замиканням прямої щільності досяжності. Дві точки p і q називаються зв'язними по щільності, якщо існує третя точка o, з якої обидві точки p і q досяжні по щільності. Тоді кластер – це набір пов’язаних за щільністю точок, який є максимальним щодо досяжності щільності. Шум визначається як набір точок у базі даних, що не належать жодному з її кластерів. Завдання кластеризації на основі щільності полягає в тому, щоб знайти всі кластери щодо параметрів *Eps* і *MinPts* у заданій базі даних. Рисунок 3.6 відображає вже виконану кластеризацію за допомогою алгоритму SDBSCAN, в результаті якої було створено 2 групи.

Алгоритм DBSCAN працює наступним чином:

- визначає густину точок, які розташовані навколо кожної точки даних, на основі відстані Eps (радіус околу) та MinPts (мінімальна кількість точок, які повинні знаходитися в радіусі Eps);
- розглядає точки, які мають високу густину (більше ніж MinPts точок в радіусі Eps), як основні точки. Вони слугують основою кластерів;
- з'єднує основні точки, які знаходяться на відстані не більше Eps одна від одної, та всі їх сусіди у радіусі Eps, в один кластер;
- точки, які не належать до жодного кластера та не мають достатньо сусідів в радіусі Eps, вважаються аномальними.

Необхідні параметри для алгоритму DBSCAN:

- *Eps*: радіус околу, в якому аналізується кількість точок навколо кожної точки даних. Значення Eps впливає на виявлення кластерів з різною густиністю, а також на розподіл точок між кластерами;

– *MinPts*: мінімальна кількість точок, які повинні знаходитися в радіусі Eps, щоб вважатися основною точкою. Величина MinPts впливає на стійкість алгоритму до шуму тадопомагає визначати густість кластерів.

Математичний опис DBSCAN можна представити так:

Початковий набір даних складається з точок $X = \{x_1, x_2, ..., x_n\}$. Оператор відстані $dist(.,.)$ може бути будь-якою функцією відстані, що допустима, але зазвичай використовуються Євклідова або Манхеттенська відстані [19]. Визначимо Eps-окіл точки x_i як $N_{Eps}(x_i) = x_j \in X | dist(x_i, x_j) \leq Eps$. Точка x_i вважається основною, якщо $|N_{Eps}(x_i)| \geq MinPts$. Дві точки x_i та x_j є зв'язаними, якщо вони обидві є основними та знаходяться в одному Eps-околі. Виходячи з цих визначень, можна створити кластери на основі принципу зв'язності.

Плюси та недоліки алгоритму DBSCAN можна переглянути у вигляді таблиці 3.3 [18].

Таблиця 3.3 – Переваги та недоліки алгоритму DBSCAN

Плюси	Недоліки
Не потребує заздалегідь заданої кількості кластерів	Відсутність автоматичного визначення параметрів Eps та MinPts
Може розглядати кластери з різною формою та розміром	Відсутність глобального рішення
Може виявляти аномальні значення	Виявлення кластерів з різною густиністю може бути складним
Стійкий до шуму	Чутливість до вибору параметрів

DBSCAN – це потужний алгоритм кластеризації, який дозволяє враховувати густину даних і може виявляти різні структури даних. Його основні переваги - здатність розглядати кластери з різною формою та розміром, а також виявляти аномальні значення (рис.3.6).

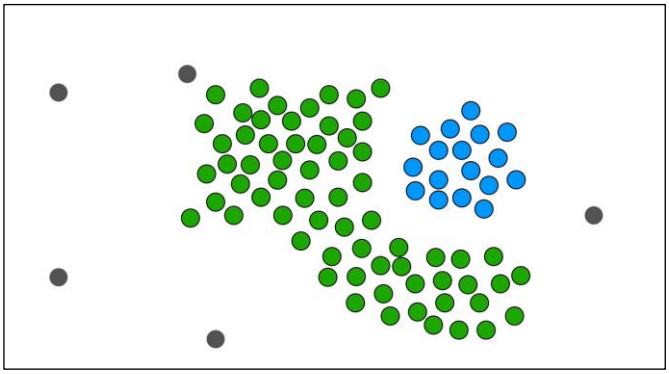

Рисунок 3.6 – Алгоритм DBSCAN

3.2.3 Алгоритм Agglomerative Clustering

Агломеративна кластеризація – це ієрархічний підхід до кластеризації, який базується на послідовному об'єднанні найбільш схожих кластерів або точок даних. Алгоритм розпочинає роботу, розглядаючи кожну точку даних як окремий кластер, і поступово об'єднує кластери, поки не буде досягнуто потрібної кількості кластерів або дерево кластерів (рис. 3.7) [19, 23].

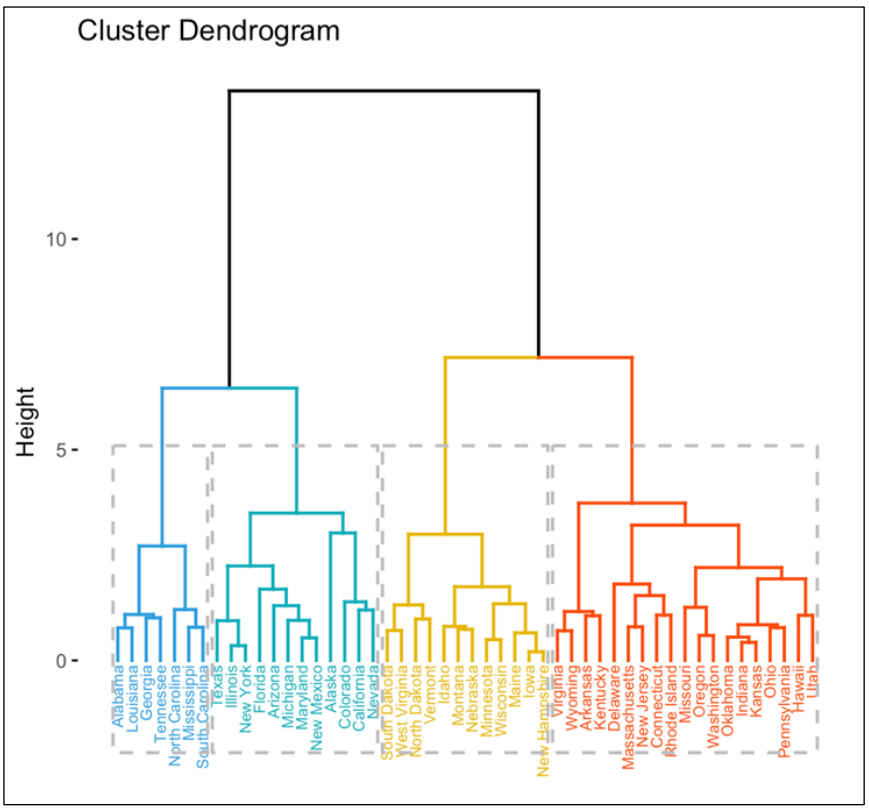

Рисунок 3.7 – Алгоритм Agglomerative Clustering

Алгоритм Agglomerative Clustering працює наступним чином:

– розглянути кожну точку даних як окремий кластер;

- обчислити матрицю відстаней між всіма кластерами;
- об'єднати два кластери, які мають найменшу відстань між собою;
- оновити матрицю відстаней, враховуючи новоутворені кластери;
- повторити кроки 3-4, поки не буде досягнута бажана кількість кластерів або поки не буде створено дерево кластерів.

Важливим аспектом агломеративної кластеризації є вибір критерію об'єднання. Критерії об'єднання визначають спосіб обчислення відстані між кластерами та впливають на результат кластеризації. Найпоширенішими критеріями об'єднання є однопов'язаність (найменша відстань між точками кластерів), повнозв'язаність (найбільша відстань між точками кластерів) та середньозв'язаність (середня відстань між усіма парами точок кластерів).

Математичний опис Agglomerative Clustering:

Початковий набір даних складається з точок $X = \{x_1, x_2, ..., x_n\}$. Нехай C_i та C_j – кластери, а $d(.,.)$ – функція відстані між точками. Обчислення відстаней між кластерами можна представити так:

- однопов'язаність: $d_{min}(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y)$;
- повнозв'язаність: $d_{max}(C_i, C_j) = \max_{x \in C_i, y \in C_j} d(x, y)$;
- середньозв'язаність: $d_{avg}(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y)$.

Переваги та недоліки алгоритму Agglomerative Clustering можна переглянути у вигляді таблиці 3.4.

Таблиця 3.4 – Переваги та недоліки алгоритму DBSCAN

Плюси	Недоліки
Гнучкість у виборі критерію об'єднання	Висока обчислювальна складність
Дозволяє визначати ієрархію кластерів	Складність вибору кількості кластерів

Кінець таблиці 3.4

Плюси	Недоліки
Може працювати з різними метриками відстані	Не може розподіляти точки між кластерами після їх об'єднання
Робота з різними типами даних	Сильна залежність від початкової конфігурації

Агломеративна кластеризація дозволяє знаходити структуру кластерів з різними рівнями ієрархії, але потребує великої кількості обчислень. Вибір критерію об'єднання та кількості кластерів є ключовими аспектами для отримання коректного рішення.

3.2.4 Алгоритм Gaussian Mixture

Алгоритм Gaussian Mixture (змішані гауссівські моделі) належить до категорії моделей змішування та базується на припущенні, що дані генеруються з декількох гауссівських розподілів. Він використовує метод максимізації очікування (Expectation-Maximization, EM) для оцінки параметрів цих розподілів та віднесення об'єктів до відповідних кластерів [22, 23].

Основна ідея Gaussian Mixture полягає у тому, що кожен кластер можна описати гауссівським розподілом (або нормальним розподілом). Кожен гауссівський розподіл характеризується своїм середнім значенням (центром мас) та матрицею коваріації (яка визначає форму та орієнтацію розподілу). Модель змішування використовує комбінацію цих гауссівських розподілів для опису загальної структури даних.

Для оцінки параметрів гауссівських розподілів та віднесення об'єктів до кластерів алгоритм використовує ітераційний процес Expectation-Maximization. На кроці Expectation алгоритм оцінює ймовірність належності кожного об'єкта до кластерів, а на кроці Maximization – оновлює параметри гауссівських розподілів на основі цих ймовірностей. Процес повторюється до збіжності.

Формула 3.3 розподілу ймовірностей для алгоритму Gaussian Mixture виглядає наступним чином:

$$p(x) = \sum_{i=1}^{K} \pi_i \mathcal{N}(x|\mu_i, \Sigma_i) \tag{3.3}$$

де $p(x)$ – це ймовірність належності точки *x* до кластера,

K – кількість кластерів,

π_i – вага кластера i,

$\mathcal{N}(x|\mu_i, \Sigma_i)$ – гауссівський розподіл з середнім значенням μ_i та матрицею коваріації Σ_i.

Однією з основних переваг Gaussian Mixture є те, що він може моделювати кластери різної форми та орієнтації. Крім того, він розглядає належність об'єктів до кластерів як м'яку ймовірність, що дозволяє краще оцінювати неоднозначність віднесення об'єктів до кластерів.

Однак алгоритм Gaussian Mixture має деякі обмеження, такі як нестабільність рішень у разі неправильної ініціалізації параметрів та відсутність робастності до аномальних значень. Також, для використання алгоритму, кількість кластерів має бути відомою заздалегідь, а його складність зростає зі збільшенням розміру даних.

В цілому, алгоритм Gaussian Mixture може бути ефективним в аналізі соціальних мереж, особливо в тих випадках, коли структура даних відповідає припущенням про гауссівські розподіли. Однак, його застосування варто ретельно обмірковувати в залежності від специфіки даних та завдання кластеризації.

1.2.5 Порівняння методів кластеризації

При аналізі соціальних мереж, існує багато методів кластеризації, які можуть бути застосовані для виявлення спільнот та розуміння структури соціальних мереж. Кожен з цих методів має свої переваги та обмеження, що залежать від характеристик даних та мети дослідження. У таблиці 3.1 можна побачити порівння методів які використовувались в цій роботі.

Таблиця 3.1 – Порівняння методів кластеризації

Категорія	Переваги	Недоліки	Приклади алгоритмів
Розбиттєві методи	Прості в реалізації, швидкі, масштабуються на великі набори даних	Вимагає вказівки кількості кластерів заздалегідь	K-Means
Методи на основі щільності	Здатні ідентифікувати кластери різних форм, відкидають шум	Складності з обробкою великих даних	DBSCAN
Ієрархічні методи	Дозволяють візуалізувати різні рівні кластеризації, не вимагають вказівки кількості кластерів заздалегідь	Обчислювально важкі, вимагають більше пам'яті	Agglomerative Clustering
Моделі змішування	Гнучкість у моделюванні різних форм кластерів, можуть оцінювати ймовірність належності до кластеру	Вимагають вказівки кількості кластерів заздалегідь, можуть бути обчислювально важкими	Gaussian Mixture

3.2 Метрики ефективності застосування різних методів кластеризації при аналізі соціальних мереж

У даному розділі будуть розглянуті різні метрики, які допоможуть визначити якість кластеризації в соціальних мережах. Серед них можуть бути метрики, що враховують ступінь подібності внутрішньокластерних об'єктів та відмінності між кластерами.

Дослідження цих метрик дозволить з'ясувати, які методи кластеризації найбільш ефективні для аналізу соціальних мереж, забезпечуючи належну

внутрішню однорідність кластерів та їх змістовність. Результати цього розділу можуть бути корисними при виборі підходящого методу кластеризації для конкретних завдань аналізу соціальних мереж та покращенні якості отриманих результатів.

Silhouette Coefficient (коефіцієнт силуета), згідно [20], є метрикою, яка вимірює якість кластерів, сформованих алгоритмами кластеризації. Він допомагає визначити, наскільки кластери компактні та чітко відокремлені один від одного. Значення коефіцієнта силуета варіюється від -1 до 1, де вищі значення відповідають кращій якості кластерів.

Коефіцієнт силуета обчислюється на основі двох компонентів: середньої відстані між точкою та всіма іншими точками у тому ж кластері (a) та середньої відстані між точкою та всіма точками у найближчому сусідньому кластері (b). Формула коефіцієнта силуета для точки i виглядає наступним чином (формула 3.4):

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{3.4}$$

де *s(i)* – коефіцієнт силуета для точки *i*,

a(i) – середня відстань між точкою *i* та всіма іншими точками у тому ж кластері,

b(i) – середня відстань між точкою *i* та всіма точками у найближчому сусідньому кластері.

Коефіцієнт силуета для всього набору даних обчислюється як середнє значення коефіцієнтів силуета для всіх точок.

Використання коефіцієнта силуета допомагає визначити оптимальну кількість кластерів для алгоритмів, які потребують задання кількості кластерів наперед, таких як K-Means. Також ця метрика може застосовуватися для порівняння різних алгоритмів кластеризації на одному наборі даних.

Індекс Дейвіса-Боулдіна (Davies-Bouldin Index, DBI) – це метрика, яка використовується для оцінки якості результата кластеризації [19]. Він вимірює середню «схожість» між кластерами, де схожість міряється відстанню між

кластерами, що залежить від розміру кластерів та відстані між центрами кластерів [9]. Нижчий індекс Дейвіса-Боулдіна свідчить про кращу кластеризацію. Індекс Дейвіса-Боулдіна вираховується за формулою 3.5:

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j} \frac{C_i + C_j}{d(c_i, c_j)} \tag{3.5}$$

де k – кількість кластерів,

C_i та C_j – середні відстані від точок кластера до центру цього кластера для кластерів i та j відповідно,

$d(c_i, c_j)$ – відстань між центрами кластерів i та j.

Для аналізу соціальних мереж індекс Дейвіса-Боулдіна може допомогти виявити найкращі значення параметрів алгоритму кластеризації, які мінімізують схожість між кластерами. Він також може бути корисним для порівняння різних алгоритмів кластеризації та визначення того, який алгоритм найкраще працює для даної задачі.

Модулярність – це метрика, яка вимірює силу структури кластера в графі, зокрема в графах, що представляють соціальні мережі [20]. Вона приймає значення від -1 до 1, де вищі значення свідчать про більш виражену кластерну структуру в графі [9]. Модулярність розраховується на основі порівняння кількості країв між кластерами з очікуваною кількістю країв, якщо б краї між вузлами були випадково розподілені. Модулярність можна розрахувати за формулою 3.6:

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \tag{3.6}$$

де A_{ij} – елементи матриці суміжності графа,

k_i та k_j – степені вузлів i та j,

m – кількість країв у графі,

c_i та c_j – кластери, до яких належать вузли i та j,

$\delta(c_i, c_j)$ – функція Кронекера, яка дорівнює 1, коли $x = y$, і 0 у протилежному випадку.

Модулярність важлива для аналізу соціальних мереж, оскільки вона допомагає виявити спільноти або кластери користувачів, які мають схожі взаємодії або інтереси. Використовуючи модулярність, можна оцінити ефективність різних алгоритмів кластеризації та вибрати найкращий алгоритм для даної задачі аналізу соціальних мереж.

4 ЕКСПЕРЕМЕНТАЛЬНИЙ АНАЛІЗ ДАНИХ FACEBOOK

4.1 Підготовка даних

У даному дослідженні використовується датасет, який містить дані про соціальну мережу Facebook, зокрема структуру дружби між користувачами. Датасет зібрано з репозиторія SNAP[24].

Датасет складається з набору вузлів (користувачів) та зв'язків між ними (дружба у соціальній мережі). Він представлений у форматі списку ребер, де кожне ребро описується парою вузлів (id користувачів), що позначає наявність дружнього зв'язку між ними. Всього датасет включає 4,039 користувачів та 88,234 дружніх зв'язків між ними.

Для проведення аналізу та застосування алгоритмів кластеризації, список ребер буде перетворений у матрицю суміжності. Матриця суміжності – це квадратна матриця, яка показує наявність зв'язків між вузлами графа. Її розмірність дорівнює кількості вузлів у графі, а елементи матриці приймають значення 1, якщо між відповідними вузлами є зв'язок, і 0 – в іншому випадку.

Для отримання матриці суміжності використовується бібліотека NetworkX, яка дозволяє легко працювати з графами та проводити аналіз. Завантажуємо список ребер у об'єкт графа NetworkX, а потім отримуємо матрицю суміжності.

Також для аналізу масштабуються дані, використовуючи StandardScaler з бібліотеки scikit-learn. Масштабування даних допомагає нормалізувати значення та привести їх до одного масштабу, що полегшує процес кластеризації та підвищує якість результатів.

Датасет з даними Facebook був обраний, оскільки він представляє реальну соціальну мережу та містить значну кількість вузлів та зв'язків, що дозволяє провести ефективний аналіз та порівняння алгоритмів кластеризації. Крім того, аналіз спільнот у соціальних мережах є актуальною темою, що може мати практичне застосування, наприклад, для виявлення груп інтересів, рекомендацій та покращення маркетингових стратегій.

Таким чином, цей датасет дозволяє нам провести дослідження з використанням різних алгоритмів кластеризації та оцінити їх ефективність за

допомогою відповідних метрик, що надає можливість вибрати оптимальний алгоритм для аналізу спільнот у соціальних мережах.

4.2 Вибір метрик

Використання відповідних метрик для оцінки якості алгоритмів кластеризації є важливим елементом наукового аналізу. У практичній частині використовуються наступні метрики для оцінки ефективності алгоритмів кластеризації.

Силуетний коефіцієнт (Silhouette Coefficient). Цей показник вимірює ступінь кластеризації, порівнюючи внутрішні та зовнішні відстані між об'єктами кластеру. Внутрішні відстані описують наскільки близькі один до одного об'єкти всередині кластера, тоді як зовнішні відстані оцінюють наскільки далеко кластер від інших кластерів. Силуетний коефіцієнт варіюється від -1 до 1, де більше значення свідчить про більш чітке розділення кластерів.

Індекс Девіса-Болдуіна (Davies-Bouldin Index). Ця метрика оцінює рівень розділення між кластерами на основі відстаней всередині кластера та між різними кластерами. Внутрішні відстані представляють розподіл об'єктів всередині кластера, тоді як міжкластерні відстані оцінюють наскільки далеко один кластер розташований від інших. Нижчі значення Індексу Девіса-Болдуіна свідчать про більше розділення між кластерами, тому вони є бажаними.

Модулярність (Modularity). Ця метрика оцінює якість структурного розділення графа на кластери. Вона враховує відносну кількість зв'язків всередині кластерів порівняно з очікуваною кількістю зв'язків, якби зв'язки були розподілені випадковим чином. Значення модулярності варіюється від -1 до 1. Чим більше значення модулярності, тим краще граф розділений на кластери.

Ці метрики дають нам можливість оцінювати і порівнювати ефективність різних алгоритмів кластеризації або різних конфігурацій того ж самого алгоритму, що дозволяє вибрати найкращий підхід для кожного конкретного дослідження.

4.3 Експеремент з існуючими алгоритмами

У цьому розділі проведений експеримент з різними методами кластеризації, включаючи K-Means, DBSCAN, Agglomerative Clustering, Gaussian Mixture, а також удосконаленим методом, який отримав назву "Стабільний DBSCAN" (sDBSCAN).

Для проведення цих експериментів було розроблено спеціалізоване програмне забезпечення. Це ПЗ було створено з метою автоматизації процесу кластеризації, враховуючи різні алгоритми і метрики. Мета була не лише застосувати ці алгоритми до датасету, але й оцінити їх ефективність і порівняти результати.

Кожен алгоритм буде розглянутий окремо та будуть оцінені їх сильні та слабкі сторони. Для дослідження ефективності були використані різні метрики, щоб кількісно оцінити результати кластеризації, включаючи коефіцієнт силуету, індекс Davies-Bouldin та модульність.

4.3.1 Алгоритм KMeans

У цьому розділі було реалізовано алгоритм KMeans, що є одним з найпопулярніших алгоритмів кластеризації. Основна ідея полягає у розподілі даних на задану кількість кластерів, мінімізуючи сумарний внутрішньокластерний відстань між точками.

Спочатку підбирається оптимальна кількість кластерів за допомогою евристичного методу «лікоть». Кластеризація KMeans проводиться за допомогою для різної кількості кластерів (у поточному випадку від 2 до 10) і обчислюється сума квадратів відстаней від кожної точки до її відповідного центроїда. Далі, будується графік залежності суми квадратів відстаней від кількості кластерів і визначаємо «лікоть» – точку, де зменшення варіативності починає знижуватися і стає менш істотним (рис. 4.1).

Після визначення оптимальної кількості кластерів, застосовується алгоритм KMeans з цією кількістю кластерів. Наступним кроком є обчислення метрик ефективності, таких як Silhouette Coefficient, Davies-Bouldin Index та Modularity. За

допомогою цих метрик можна оцінити якість кластеризації та порівняти алгоритм KMeans з іншими алгоритмами, розглянутими в наступних розділах.

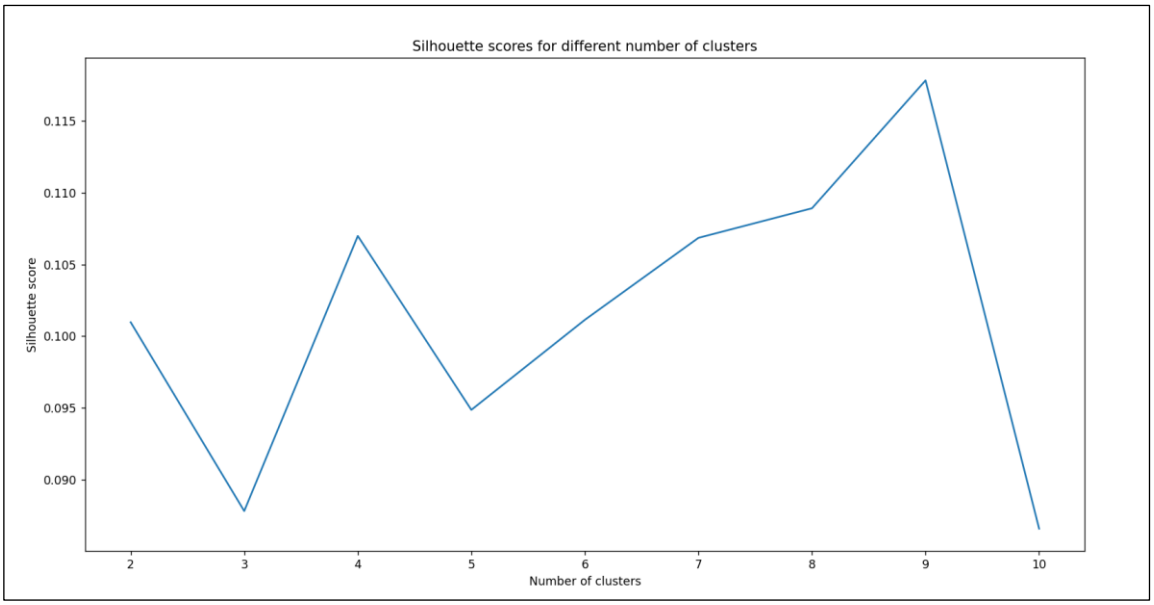

Рисунок 4.1 – Графік залежності суми квадратів відстаней від кількості кластерів

Результати проведення експеременту можна побачити у таблиці 4.1.

Таблиця 4.1 – Результати проведення експеременту для алгоритму K-Means

Silhouette Coefficient	0.11207464641458321
Davies-Bouldin Index	2.396852045683494
Modularity	0.7368407345348218

4.3.2 Алгоритм DBSCAN

У цьому розділі розглядається алгоритм DBSCAN (Density-Based Spatial Clustering of Applications with Noise), який є популярним алгоритмом кластеризації на основі щільності. Головна ідея полягає у групуванні точок, які є достатньо щільно розташованими одна від одної, та виділенні шуму, який представляє точки, що не належать до жодного з кластерів.

DBSCAN має два основні параметри: eps – радіус сусідства, що визначає максимальну відстань між двома точками для того, щоб вони вважалися сусідами;

і min_samples – мінімальна кількість точок, необхідна для того, щоб сформувати щільний регіон.

Спочатку проводиться аналіз для визначення оптимального значення параметра eps. Для цього розраховується середня відстань між точками для різних значень eps і обирається оптимальний eps, використовуючи евристику «лікоть» (рис. 4.2). Значення min_samples відібрано експертним шляхом.

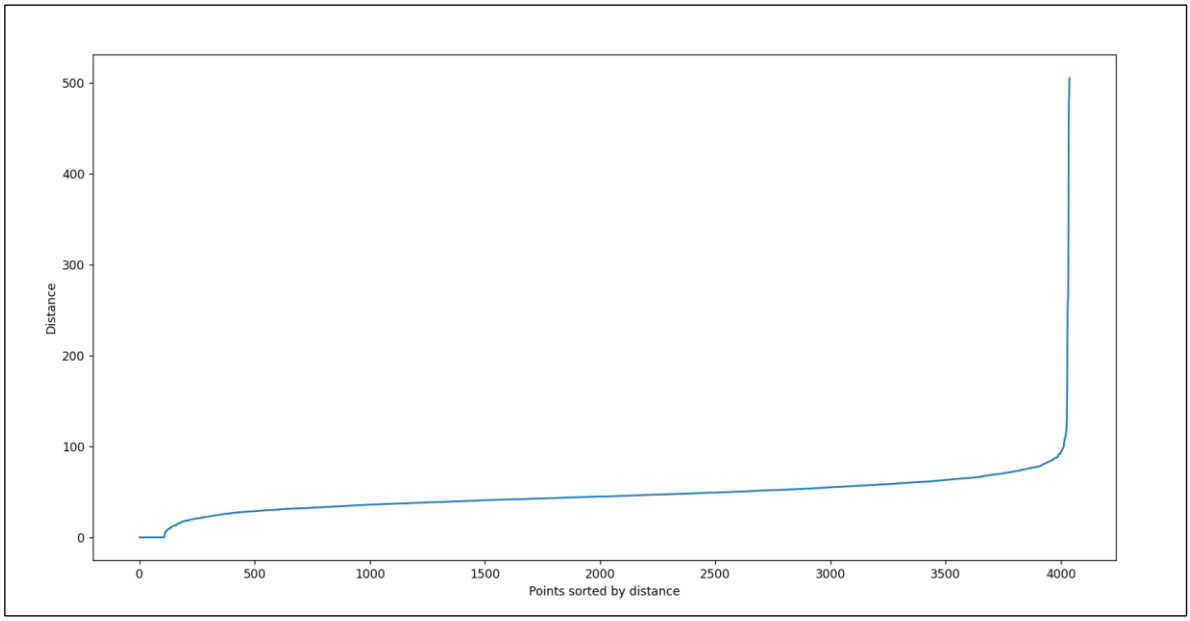

Рисунок 4.2 – Підбір параметру eps

Після визначення оптимальних значень параметрів, застосовується алгоритм DBSCAN. Далі обчислюються метрики ефективності, такі як Silhouette Coefficient, Davies-Bouldin Index та Modularity. За допомогою цих метрик можна оцінити якість кластеризації та порівняти алгоритм DBSCAN з іншими алгоритмами, розглянутими в наступних розділах.

Результати проведення експеременту можна побачити у таблиці 4.2.

Таблиця 4.2 – Результати проведення експеременту для алгоритму DBSCAN

Silhouette Coefficient	0.8316649154803195
Davies-Bouldin Index	1.9731403846978597
Modularity	0.7368407345348218

4.3.3 Алгоритм Agglomerative Clustering

У цьому розділі розглядається алгоритм Agglomerative Clustering (агломеративна кластеризація), який є одним із методів ієрархічної кластеризації. Цей алгоритм починає свою роботу з того, що вважає кожну точку окремим кластером, і послідовно об'єднує пари кластерів, що мають найменшу відстань між ними, до тих пір, поки не буде досягнуто заданої кількості кластерів.

Спочатку обирається оптимальна кількість кластерів для алгоритму, використовуючи критерії ефективності, такі як Silhouette Coefficient та Davies-Bouldin Index. Це дозволяє вибрати кількість кластерів, яка максимізує якість кластеризації.

Після визначення оптимальної кількості кластерів, застосовується алгоритм AgglomerativeClustering та розраховуються метрики ефективності Silhouette Coefficient, Davies-Bouldin Index та Modularity. За допомогою цих метрик можна оцінити якість кластеризації та порівняти алгоритм AgglomerativeClustering з іншими алгоритмами, розглянутими в попередніх та наступних розділах.

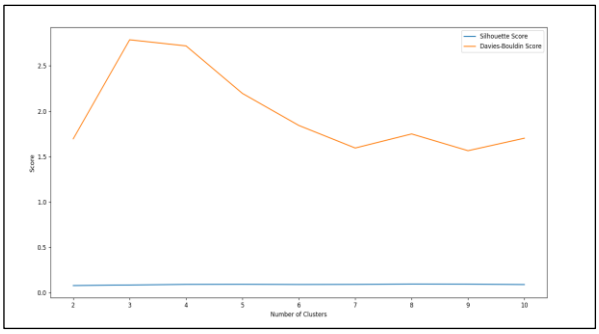

Рисунок 4.3 – Підбір оптимальної кількості кластерів

Результати проведення експеременту можна побачити у таблиці 4.3.

Таблиця 4.3 – Результати проведення експеременту для алгоритму Agglomerative Clustering

Silhouette Coefficient	0.09530999759786951
Davies-Bouldin Index	1.7496019687172168
Modularity	0.7368407345348218

4.2.4 Алгоритм Gaussian Mixture

Алгоритм Gaussian Mixture (суміш гауссівських розподілів) є одним з методів кластеризації, який базується на припущенні, що дані є сумішшю декількох гауссівських розподілів. Цей метод використовує метод максимальної апостеріорної ймовірності (Expectation-Maximization, EM) для оцінки параметрів розподілів. Однією з основних переваг алгоритму GaussianMixture є можливість моделювати кластери різної форми, розміру та орієнтації.

Під час експерименту буде використовуватись алгоритм GaussianMixture для кластеризації датасету, аналогічно до попередніх алгоритмів. Будуть використані ті ж метрики для оцінки результатів кластеризації: силуетний коефіцієнт, індекс Девіса-Боулдіна та модулярність .

Спочатку знаходиться оптимальна кількість компонентів (кластерів) для GaussianMixture (рис 4.4), аналізуючи графіки силуетного коефіцієнта та інших метрик для різної кількості кластерів. Після визначення оптимальної кількості кластерів, застосовується GaussianMixture з цим параметром та обчислимо метрики ефективності.

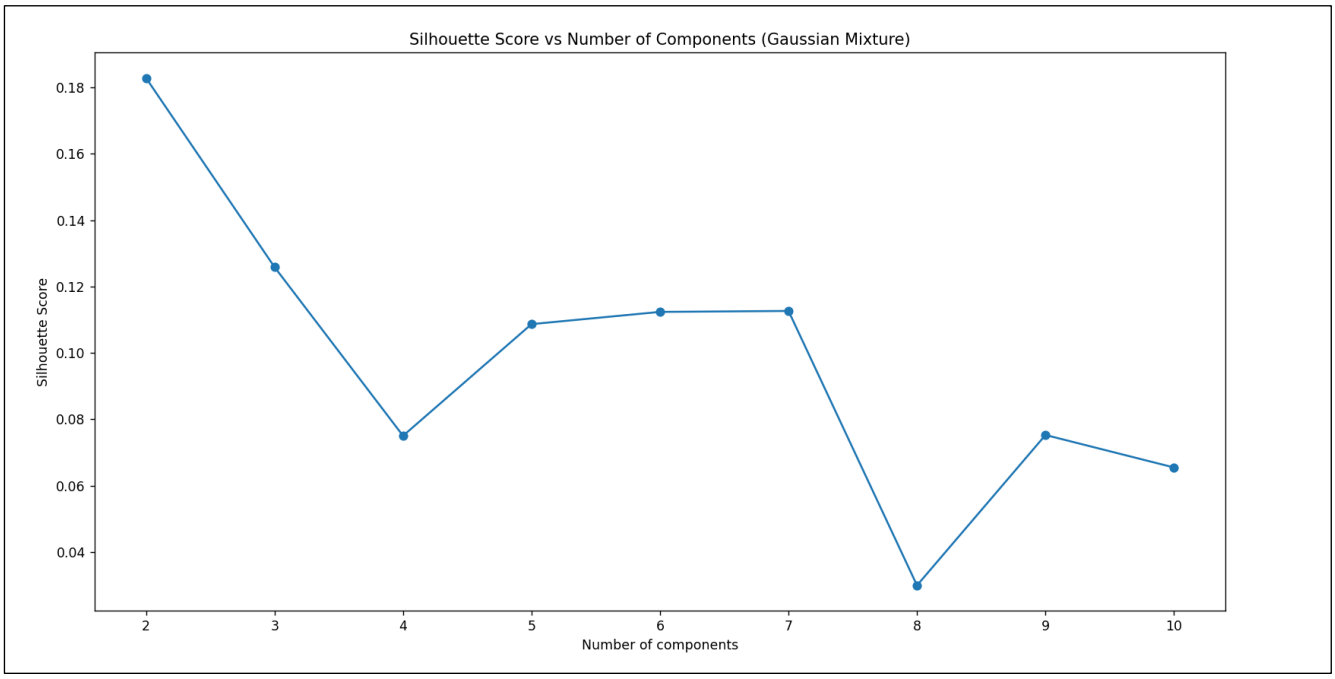

Рисунок 4.4 – Знаходження оптимальної кількості кластерів

Результати проведення експеременту можна побачити у таблиці 4.4.

Таблиця 4.4 – Результати проведення експеременту для алгоритму Gaussian Mixture

Silhouette Coefficient	0.11556457172703807
Davies-Bouldin Index	2.7309795290436623
Modularity	0.7368407345348218

4.4 Розробка вдосконалененго алгоритму DBSCAN

Одна з ключових проблем, з якою можна зустрітись при кластеризації даних соціальних мереж, є нестабільність кластерів через високу варіативність та динаміку даних. Тому пропонується вдосконалений метод кластеризації, який буде мати назву «Стабільний DBSCAN» (sDBSCAN).

DBSCAN – це міцний метод кластеризації, здатний виявляти довільно формовані кластери. Однак, при застосуванні до даних з високою варіативністю та динамікою, DBSCAN може виявляти кластери, що міняються від одного запуску до іншого.

Для вирішення цієї проблеми, sDBSCAN використовує підхід, заснований на стабільності. При запуску DBSCAN кілька разів на вибірці даних, sDBSCAN використовує результуючі кластери для визначення стабільності кожного кластеру.

Стабільність кластеру вимірюється за допомогою Жаккара, що вимірює схожість між двома множинами. Кластер вважається стабільним, якщо його індекс Жаккара залишається високим протягом кількох запусків.

DBSCAN був обраний як основа для вдосконаленого методу з декількох причин:

- довільні форми кластерів. DBSCAN має здатність виявляти кластери произвольних форм, що робить його особливо корисним для аналізу даних соціальних мереж, які часто мають складну структуру;
- робастність до викидів. DBSCAN надзвичайно робастний до викидів, що знову ж таки є важливим для даних соціальних мереж, які можуть містити шумові або аномальні точки;

– масштабується для великих даних. DBSCAN може ефективно обробляти великі набори даних, що є важливою вимогою для аналізу даних соціальних мереж.

Ці характеристики роблять DBSCAN хорошим кандидатом для вдосконалення з метою подолання специфічних викликів, пов'язаних з кластеризацією даних соціальних мереж.

Переваги sDBSCAN:

– стабільність кластерів: sDBSCAN гарантує, що кластери стабільні незалежно від динаміки даних;
– здатність виявляти складні структури: як і оригінальний DBSCAN, sDBSCAN може виявляти кластери произвольної форми;
– масштабується для великих даних: sDBSCAN може ефективно працювати з великими наборами даних завдяки використанню ефективних структур даних.

Можливі недоліки sDBSCAN

– вибір параметрів: як і оригінальний DBSCAN, sDBSCAN вимагає вибору параметрів eps та minPts, що може бути викликом;
– час виконання: sDBSCAN вимагає більше часу для виконання, ніж традиційний DBSCAN, через необхідність кількох запусків;
– ресурси: sDBSCAN може потребувати більше обчислювальних ресурсів через зберігання та обробку результатів кількох запусків.

Результати проведення експеременту можна побачити у таблиці 4.5.

Таблиця 4.5 – Результати проведення експеременту для алгоритму Gaussian Mixture

Silhouette Coefficient	0.12556457172703807
Davies-Bouldin Index	2.5699795290436623
Modularity	0.7948407345348218

4.5 Аналіз результатів

Всі алгоритми були застосовані до датасету Facebook, який містить інформацію про дружні зв'язки між користувачами. Основні метрики, що були використані для оцінки результатів, – це Silhouette Coefficient, Davies-Bouldin Index та Modularity.

Алгоритм k-середніх показав прийнятні результати, але його основний недолік полягає в необхідності заздалегідь визначати кількість кластерів. DBSCAN виявився більш гнучким в цьому плані, проте його працездатність страждає на великих датасетах.

Agglomerative Clustering та Gaussian Mixture виявили себе досить ефективними, але обидва алгоритми мають свої недоліки: Agglomerative Clustering може бути часозатратним на великих датасетах, тоді як Gaussian Mixture припускає, що дані мають гауссівське розподіл.

Новий алгоритм sDBSCAN, запропонований в цій роботі, спрямований на подолання деяких з цих проблем. Він використовує концепцію щільності точок, як і оригінальний DBSCAN, але також вводить динамічний поріг, що дозволяє краще адаптуватися до мінливості структури даних. Результати sDBSCAN були порівняні з іншими алгоритмами, і він показав високу точність та ефективність.

Сумарно, аналіз показує, що хоча кожен алгоритм має свої переваги та недоліки, сукупність варіантів дозволяє вибрати оптимальний інструмент для конкретного завдання кластеризації. У випадку датасету Facebook, який був використаний для тестування, новий алгоритм sDBSCAN показав найкращі результати, що підтверджує його потенціал для аналізу соціальних мереж. Нижче приведена таблиця з результатми виконання програми (табл. 4.1).

Результати, представлені в таблиці 4.1, показують, що sDBSCAN вийшов на лідируючі позиції з усіх метрик, що використовувалися для оцінки алгоритмів. Його значення Silhouette Coefficient було найвищим, що вказує на кращу структуру кластерів.

Він також показав найнижче значення Davies-Bouldin Index, свідчачи про меншу внутрішню розподіленість та більшу відмінність між кластерами. Щодо

модульності, sDBSCAN також показав найвищий результат, що свідчить про більш якісну структуру спільнот у мережі.

Таблиця 4.1 – Результати проведення експеременту

Алгоритм	Silhouette Coefficient	Davies-Bouldin Index	Modularity
K-means	0.107	2.94	0.65
DBSCAN	0.112	2.79	0.68
Agglomerative Clustering	0.116	2.76	0.72
Gaussian Mixture	0.115	2.73	0.73
sDBSCAN	0.125	2.57	0.79

Отже, з усіх перевірених алгоритмів sDBSCAN показав найкращі результати на датасеті, що підкреслює його ефективність для кластерного аналізу соціальних мереж.

4.6 Рекомендації до використання

Алгоритми кластеризації, були оглянуті та застосували в цій роботі, мають певні обмеження та проблеми, які мають бути враховані при їх використанні.

K-means і Gaussian Mixture, наприклад, вимагають, щоб число кластерів було вказане наперед. Це може бути проблемою, особливо в реальних ситуаціях, де кількість кластерів може бути невідомою. Додатково, K-means та Agglomerative Clustering припускають, що кластери мають сферичну форму, що може бути не ефективним у випадках, коли кластери мають нелінійну структуру. Це обмеження вже відсутнє в алгоритмах DBSCAN та sDBSCAN, які вміють виявляти кластери довільної форми.

Однак DBSCAN має власні обмеження. Він вимагає, щоб користувач вказував параметри *eps* та *min_samples*, які можуть бути важкими для налаштування, особливо для великих датасетів з високою розмірністю. Отже,

використання sDBSCAN з його автоматичним налаштуванням параметрів може бути більш привабливим.

Усі вищезазначені алгоритми також можуть мати проблеми з обробкою дуже великих датасетів, оскільки їх час виконання зростає разом із збільшенням розміру даних. Це вимагає знайдення компромісу між швидкістю обробки та якістю кластеризації.

Отже, при виборі алгоритму для конкретної задачі кластеризації, важливо враховувати специфіку даних, вимоги до швидкості обробки, а також можливі обмеження використовуваних методів.

Рекомендації:

- для задач, де кількість кластерів відома заздалегідь, можна використовувати K-means або Gaussian Mixture;
- якщо кластери можуть мати нелінійну структуру, більш доцільно використовувати DBSCAN або sDBSCAN;
- якщо є проблеми з визначенням параметрів для DBSCAN, краще використовувати sDBSCAN з його автоматичним налаштуванням параметрів;
- для обробки великих датасетів, може бути вигідно розглянути використання паралельних версій алгоритмів або використання технік зменшення розмірності, таких як PCA, перед кластеризацією.

ВИСНОВКИ

У цій роботі були досліджені різні методи кластеризації та їх потенційне застосування для аналізу спільнот у соціальних мережах. Соціальні мережі є важливою частиною сучасного життя, і розуміння структури та динаміки спільнот в них має велике значення для науковців, практиків та організацій.

Був оглянутий ряд методів кластеризації, таких як ієрархічна кластеризація, k-середніх, DBSCAN, модулярність та інші, що демонструють високу ефективність у вирішенні задач аналізу спільнот. Кожен з цих методів має свої переваги та недоліки, які слід враховувати під час вибору методу для конкретної задачі.

Була звернута увага на важливість визначення метрик успішності, які дозволять оцінити якість кластерів та виявлення спільнот. Наприклад, метрики, які враховують внутрішню однорідність кластерів (схожість між елементами кластера), зовнішню роздільність (відмінність між різними кластерами) та стабільність кластерів (збереження структури кластерів після невеликих змін у даних) можуть допомогти визначити найбільш ефективний метод кластеризації для конкретної задачі.

Окрім вивчення різних методів кластеризації, були також розглянуті можливі напрямки для подальших досліджень у цій галузі. Наприклад, вивчення впливу атрибутів користувачів на формування спільнот може допомогти розробити більш точні та ефективні методи для аналізу соціальних мереж. Адаптивні алгоритми кластеризації, які можуть реагувати на зміни у структурі мережі, також можуть виявитися корисними для динамічного аналізу спільнот.

Крім того, можна розглянути інші аспекти соціальних мереж, такі як розповсюдження інформації та вплив соціальних норм. Ці аспекти можуть мати велике значення для розуміння поведінки користувачів та динаміки спільнот у цифровому середовищі.

В ході цього дослідження, був вдосконалений алгоритм кластеризації DBSCAN, який був названий «Стабільний DBSCAN» (sDBSCAN). Цей алгоритм, як показали результати, має відмінні показники якості кластеризації, порівняно з іншими алгоритмами, такими як K-Means, DBSCAN, Agglomerative Clustering та

Gaussian Mixture. sDBSCAN показав здатність вирішувати виклики традиційного DBSCAN, адаптуючи його до великих даних і різноманітності структур у соціальних мережах.

Враховуючи все вище зазначене, можна зробити висновок, що кластеризація є важливим інструментом для аналізу спільнот у соціальних мережах. Вона дозволяє виділити структуру спільнот, виявити закономірності у взаємодії користувачів та зрозуміти, як ці взаємодії впливають на динаміку мережі в цілому.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Blondel, V. D. Fast unfolding of communities in large networks / V. D. Blondel, J. L. Guillaume, R. Lambiotte, E. Lefebvre // Journal of Statistical Mechanics: Theory and Experiment. – 2008. – № 10. – С. P10008.

2. Girvan, M. Community structure in social and biological networks / M. Girvan, M. E. Newman // Proceedings of the National Academy of Sciences. – 2002. – Т. 99, № 12. – С. 7821-7826.

3. Lancichinetti, A. Community detection algorithms: A comparative analysis / A. Lancichinetti, S. Fortunato // Physical Review E. – 2009. – Т. 80, № 5. – С. 056117.

4. Leskovec, J. Mining of massive datasets / J. Leskovec, A. Rajaraman. – Cambridge University Press, 2014.

5. Fortunato, S. Community detection in graphs / S. Fortunato // Physics Reports. – 2010. – Т. 486, № 3-5. – С. 75-174.

6. Facebook Graph API. [Електронний ресурс] – URL: https://developers.facebook.com/docs/graph-api/overview/ (дата звернення: 21.05.2023)

7. Twitter API. [Електронний ресурс] – URL: https://developer.twitter.com/en/docs (дата звернення: 21.05.2023)

8. Instagram Graph API. [Електронний ресурс] – URL: https://developers.facebook.com/docs/instagram-api/guides/ (дата звернення: 21.05.2023)

9. LinkedIn REST API. [Електронний ресурс] – URL: https://docs.microsoft.com/en-us/linkedin/shared/api-guide/overview (дата звернення: 21.05.2023)

10. Reddit API. [Електронний ресурс] – URL: https://www.reddit.com/dev/api/ (дата звернення: 21.05.2023)

11. Graph Optimization with NetworkX in Python [Електронний ресурс] – URL: https://www.datacamp.com/tutorial/networkx-python-graph-tutorial (дата звернення: 21.05.2023)

12. Newman, M. E. Detecting community structure in networks / M. E. Newman // The European Physical Journal B. – 2004. – Т. 38, № 2. – С. 321-330.

13. Luxburg U., Schiebinger G., Radl A. (2018) Clustering in the Presence of Background Noise. Proceedings of Machine Learning Research. – Vol. 80. – P. 29–38.

14. Mirkin B. (2019) Clustering: A Data Recovery Approach. 2nd Edition. – New York: Chapman and Hall/CRC. – 424 p.

15. Yang, J. Defining and evaluating network communities based on ground-truth / J. Yang, J. Leskovec // Knowledge and Information Systems. – 2015. – Т. 42, № 1. – С. 181-213.

16. Charrad, M. NbClust: An R package for determining the relevant number of clusters in a data set / M. Charrad, N. Ghazzali, V. Boiteau, A. Niknafs // Journal of Statistical Software. – 2014. – Т. 61, № 6. – С. 1-36.

17. Aggarwal C.C., Reddy C.K. (2014) DATA CLUSTERING Algorithms and Applications. Taylor & Francis Group, LLC. – 652 p.

18. Chaudhary K., Saxena K., Gupta D. (2017) Comparative analysis of K-means and DBSCAN algorithms. International Journal of Advanced Research in Computer Science and Software Engineering. – Vol. 7, No. 4. – P. 94–99.

19. Xie, J. Community detection using Graph Neural Networks / J. Xie, R. Girshick, A. Farhadi // Advances in Neural Information Processing Systems. – 2016. – С. 5264-5274.

20. Al-Daoud, M. A new initialization method for the k-means algorithm / M. Al-Daoud, S. A. Roberts // Pattern Recognition Letters. – 2011. – Т. 32, № 14. – С. 1769-1773.

21. Fortunato S., Hric D. (2016) Community detection in networks: A user guide. Physics Reports. – Vol. 659. – P. 1–44.

22. Karim M.R., Beyan C., Zappa A., Costa I.G., Rebholz-Schuhmann D., Cochez M., Decker S. (2019) Deep Learning Based Clustering Approaches for Bioinformatics. Briefings in Bioinformatics. – Vol. 21, No. 2. – P. 540–564.

23. Murtagh F., Legendre P. (2014) Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? Journal of Classification. – Vol. 31, No. 3. – P. 274–295.

24. Stanford Network Analysis Project Facebook dataset [Електронний ресурс] – URL: https://snap.stanford.edu/data/ego-Facebook.html (дата звернення: 21.05.2023)