DOI: https://doi.org/10.36910/6775-2524-0560-2025-60-14
УДК 004.91
Захарчук Наталія Геннадіївна, магістриня
https://orcid.org/0009-0007-2196-924X
Ткаченко Олексій Миколайович, к. т. н., доцент
https://orcid.org/0000-0002-9514-516X
Київський національний університет імені Тараса Шевченка, м. Київ, Україна

СИСТЕМА АНАЛІЗУ ГРУП КОРИСТУВАЧІВ СОЦІАЛЬНИХ МЕРЕЖ НА ОСНОВІ ГРАФОВИХ БАЗ ДАНИХ

Захарчук Н.Г., Ткаченко О.М. Система аналізу груп користувачів соціальних мереж на основі графових баз даних. У статті розглянуто підходи та інструменти аналізу соціальних мереж (СМ) з використанням графових баз даних (БД). Проведено дослідження методів аналізу, здійснено програмну реалізацію виявлення та оцінювання характеристик, у т.ч. динамічних, окремих соціальних спільнот. Розроблений інструментарій дозволяє: виявляти потенційні ботоферми (групи акаунтів з високою щільністю внутрішніх зв'язків і майже відсутніми зовнішніми контактами, синхронізацією публікацій високої подібності); виявляти прихильників та опонентів певного наративу, ідеї, партії чи особи (кластери акаунтів, що поширюють позитивні/негативні повідомлення про об'єкт, порівняння середнього поляритету і топ ключових слів у кожному кластері, визначення ключових впливових вузлів у кожній групі); оцінювати структурну інтеграцію та згуртованість спільнот – розмір кластерів, коефіцієнт кластеризації, щільність зв'язків, виявлення мостів і точок зламу для оцінки стійкості інформаційного поширення. Отримані результати підтверджують актуальність застосування графових БД, зокрема, Neo4j, у задачах аналізу складних соціальних структур. Розроблене ПЗ можна адаптувати до інших тем, джерел або платформ та розширити функціонал для різних типів соціальних мереж.
Ключові слова: соціальна мережа, аналіз, кластеризація, центральність, щільність, графова база даних

Zakharchuk N., Tkachenko O. A system for analyzing groups of social network users based on graph databases. The article examines approaches and tools for social network analysis using graph databases. The study explores methods for analysis and implements the detection and assessment of characteristics of social communities, including its dynamic. The developed toolkit enables: detecting potential bot farms (groups of accounts with high internal connection density, almost no external contacts, and synchronized publication of highly similar content); identifying supporters and opponents of a specific narrative, idea, party, or person (clusters of accounts that disseminate positive/negative messages about the target, comparing the average polarity and top keywords in each cluster, and identifying key influential nodes in each group); assessing the structural integration and cohesion of communities (cluster size, clustering coefficient, connection density, detection of bridges and break points to evaluate the resilience of information dissemination). The obtained results confirm the relevance of using graph databases, particularly Neo4j, for the analysis of complex social structures. The developed software can be adapted to other topics, sources, or platforms, and its functionality can be extended for various types of social networks.
Keywords: social network, analysis, clustering, centrality, density, graph database

Постановка наукової проблеми. Стрімке зростання популярності Інтернет-сервісів, які відносять до Web 2.0, особливо соціальні мережі (СМ), обумовило зміну парадигми поширення і сприйняття інформації, що передбачає розширення спектру інформаційних джерел та можливостей їх вибору. Разом з тим, з'явилися нові виклики, пов'язані з пропагандою, інформаційними війнами, іншими видами обману. Так, на всесвітньому економічному форумі у 2024 р. дезінформація визнана одним з глобальних ризиків [1]. Протидія таким загрозам передбачає комплекс заходів – від організаційних та освітніх на рівні держави до науково-технічних. Автоматизований аналіз груп користувачів СМ передбачає використання засобів виявлення таких груп, відслідковування їх динаміки, взаємодії з різними категоріями користувачів та іншими групами, аналіз контенту, емоційного впливу тощо.

З огляду на зазначене, обумовлюється актуальність розробки методів та програмних рішень автоматизованого аналізу груп користувачів СМ. Значне зростання популярності інтелектуальних підходів, таких як інтелектуальний аналіз даних та машинне навчання, методологічний інструментарій окресленої проблематики відкриває нові можливості для дослідників та розробників. Використання графових баз даних (БД) має ряд переваг при дослідженні СМ, наприклад, швидка обробка складних запитів при аналіз зв'язків між вузлами, що обумовлює доцільність їх використання при розробці вказаних систем.

Варто зауважити, що засоби аналізу поведінки користувачів у СМ можуть мати й інші прикладні застосування, наприклад, у маркетингу, протидії шахрайству чи розвитку спільнот різного спрямування.

Аналіз останніх досліджень і публікацій. Методам аналізу СМ присвячено ряд публікацій. Так, у [2] представлено огляд моделей СМ з огляду на взаємозалежність соціальних одиниць

(користувачів). Там же розглянуто моделі на графах, у т.ч. випадкових, а також їхню структурну еквівалентність. У [3] подано огляд та порівняння популярних методів та інструментів аналізу СМ, зокрема, методів аналізу графів та методів кластеризації, можливості застосування у різних сферах, наприклад, соціології чи біології. У статті [4] досліджено методи кластеризації графів СМ, зокрема, методи на основі модулярності графа, розмітці графа та випадкових блукань. Розглянуто використання кластеризації для побудови рекомендаційних систем, можливості графової БД Neo4j та алгоритмів з її бібліотеки, такі як Louvain, Label Propagation та Triangle Counting, які можуть спростити виявлення спільнот. У [5] проведено огляд методів виокремлення спільнот у СМ з атрибутами вузлів. Зокрема, описано підходи до виявлення спільнот з урахуванням додаткових властивостей користувачів (наприклад, їхніх інтересів чи уподобань). Подано класифікацію методів за часом інтеграції атрибутивної інформації з топологією мережі, проаналізовано їх ефективність на реальних даних. Ряд публікацій присвячені аналізу СМ з використанням методів машинного навчання, наприклад, у [6] подано загальну модель виявлення спільнот з використанням глибинного навчання. Запропоновано GCN-модель (Graph Convolutional Network), яка поєднує структуру соціального графа і зміст комунікацій для підвищення якості виявлення спільнот. У [7] проаналізовано методи виявлення типів соціальних спільнот та методи машинного навчання для їх пошуку. З'ясовано, що найпоширеніші підходи – некеровані (кластеризація), а з 2020 р. суттєво зросло застосування методів глибинного навчання для великих мереж. Найпопулярнішою метрикою оцінки якості виділених спільнот є Normalized Mutual Information.

Ряд публікацій присвячено виявленню спільнот та їх динаміці. Так, у [8] зроблено огляд підходів виявлення спільнот у динамічних СМ. Розглянуто задачу пошуку та еволюцію спільнот у мережах, які змінюються в часі. У деяких публікаціях [9] досліджено характеристики і динаміку спільнот покоління т.з. "зумерів". На датасеті взаємодій користувачів проаналізовано стабільність модулярності мережі (характеризує чіткість спільнот) у часі. Показано, що модулярність мережі є досить стабільною характеристикою, оскільки користувачі схильні до збереження кола своїх контактів, тому структура спільнот мало змінюється. Крім того, початковий рівень модулярності виявився предиктором майбутніх характеристик як всієї мережі, так і его-мереж окремих користувачів, що свідчить про важливість раннього структурування груп у життєвому циклі мережі. Робота [10] присвячена аналізу структури і динаміки взаємодій у спільноті Reddit. Автори змоделювали дискусійні треди на зростаючій спільноті і дослідили еволюцію їх властивостей. Вони виявили, що глобальний коефіцієнт кластеризації є досить малий, а середня довжина шляху між користувачами зростає з часом. Це свідчить, що обговорення на Reddit здебільшого відбуваються у формі коротких двосторонніх діалогів. Показано, що внутрішні правила спільноти впливають на саму структуру діалогів, тобто еволюція мережі треду фактично складається з двох підграфів, що зростають з різною швидкістю через особливості взаємодій, визначені модерацією конкретної спільноти. Проблемі виявлення поширювачів дезінформації на прикладі СМ X (колишній Twitter) присвячено дослідження [11], де досліджено розповсюдження дезінформації про COVID-19 із застосуванням відповідних метрик аналізу СМ. Автори запропонували архітектуру системи, що включає модулі збору даних, побудови графу користувачів, виявлення центральностей, виділення спільнот та аналізу поширення фейків. На експериментальному наборі повідомлень про пандемію підхід виявив "ненадійні" акаунти, залучені до дезінформації, та показав можливість оцінки їх впливу на інших.

Дослідженню ефективності графових БД у СМ, зокрема, Neo4j, TigerGraph, JanusGraph, Nebula, присвячена стаття [12]. За результатами випробувань показано найвищу продуктивність Neo4j за всіма метриками.

Метою роботи є аналіз сучасних підходів, інструментів аналізу та розробки програмних систем виявлення груп користувачів, їх структури, зв'язків і динаміки за окремими параметрами на основі реальних даних з використанням графових БД та інтелектуальних методів.

Виклад основного матеріалу дослідження. Результати дослідження [12] обумовили вибір обрати графової БД для дослідження і розробки. Однією з ключових характеристик графових БД є висока продуктивність при роботі зі складними запитами, які включають глибокі відносини між даними. Це досягається зберіганням зв'язків між вузлами разом з самими даними, що в результаті знижує витрати на їх визначення під час виконання запитів. Для роботи з графовими БД використовуються спеціалізовані мови запитів, такі як: Cypher, Gremlin, а також SQL-подібні мови для деяких графових систем. Ці мови дозволяють формулювати складні запити для пошуку

шаблонів, виконувати аналіз зв'язків та робити висновки. Узагальнене порівняння графових БД наведено у таблиці 1.

Таблиця 1. Порівняння популярних графових БД

Особливість	**Neo4j**	**Apache TinkerPop**	**OrientDB**	**ArangoDB**
Модель	Графова СУБД	Фреймворк для обчислень на графах	Мульти-модель (Граф/Документ)	Мульти-модель (Граф/Документ/Ключ-Значення)
Мова запитів	Cypher	Gremlin	SQL-подібна, Gremlin	AQL
Масштабо-ваність	Висока з підтримкою кластеризації	Залежить від БД, на якій базується	Висока з розподіленою конфігурацією	Висока з кластеризацією та шардуванням
Випадки використання	СМ, системи рекомендацій, виявлення шахрайства	Аналітика графів, операції обходу, підтримка між платформами	Мульти-модельні додатки, аналітика в реальному часі	Складні запити, мульти-модельні додатки

Графові БД зручно використовувати для аналізу профілів користувачів СМ, якщо позначити користувачів як вершини графу, а зв'язки між ними — як ребра.

Для розробки системи аналізу груп у СМ було обрано мову програмування Python та БД Neo4j. Однією з ключових переваг Neo4j є мова запитів Cypher, яка є спеціалізованою та зручною для роботи з графовими структурами. Це дозволяє легко формулювати запити для складних аналітичних задач, таких як визначення найкоротших шляхів, пошук шаблонів зв'язків та агрегація даних на основі графових структур.

Аналіз поведінки груп у СМ фокусується на вивченні того, як групи осіб формуються, взаємодіють, розвиваються, приймають рішення або впливають одна на одну. Методи аналізу СМ передбачають використання теорії графів для кількісного вимірювання і моделювання соціальних структур. Аналіз СМ використовує ряд фундаментальних понять та метрик, які допомагають кількісно оцінити структуру і динаміку мереж. Ці поняття та метрики забезпечують розуміння взаємодій та взаємозв'язків між елементами мережі. Розглянемо основні з них:

1. *Вузли та ребра*. Вузли – окремі елементи мережі, які можуть представляти людей, організації, об'єкти або концепції. Ребра – зв'язки між вузлами, які можуть бути напрямними або ненапрямними і часто містять інформацію про тип та силу взаємодії.

2. *Степінь вузлів* – кількість ребер, що з'єднують вузол з іншими вузлами. В аналізі СМ високий ступінь вузла часто вказує на високу соціальну активність особи або організації.

3. *Центральності*. *За* степенем: вимірює впливовість вузла на основі кількості його прямих зв'язків. Близькісна: визначає, наскільки швидко вузол може досягти всіх інших вузлів у мережі. Вузол з високою близькісною центральністю має короткі шляхи до всіх інших вузлів.

4. *Щільність мережі* відображає загальну зв'язаність мережі, визначаючи співвідношення фактичної кількості ребер до максимально можливої кількості ребер між вузлами.

Алгоритми виявлення спільнот, такі як Girvan-Newman або Лувена, зосереджуються на ідентифікації груп вузлів, які мають високу щільність внутрішніх зв'язків. Ці групи часто відображають спільноти або кластери користувачів зі схожими інтересами або взаємодіями. Результати таких алгоритмів можуть використовуватись для таргетингу аудиторії у маркетингових кампаніях або для аналізу соціальних структур.

```python
def community_detection(graph):
    communities = []
    for component in nx.connected_components(graph):
        communities.append(component)
    return communities
```

Рис. 1. Приклад реалізації алгоритму виявлення спільнот для графу на мові Python

Алгоритми виявлення спільнот (Girvan-Newman, Лувен) виділяють кластери учасників за щільністю зв'язків. PageRank і HITS (Hyperlink-Induced Topic Search) визначають значущість вузлів, базуючись на структурі входів/виходів. У СМ, такі алгоритми допомагають визначити, які користувачі мають найбільший вплив або центральність у мережі. Ці алгоритми аналізують, наскільки часто інші вузли спираються на даний вузол у мережі, що може бути індикатором впливовості особи або значущості ресурсу. Наприклад, алгоритм PageRank (рис. 2), спочатку розроблений для ранжування веб-сторінок в інтернеті, ефективно застосовується для ідентифікації ключових учасників у соціальних мережах, що впливають на розповсюдження інформації та формування громадської думки.

```python
def page_rank(graph):
    pagerank_scores = nx.pagerank(graph)
    return pagerank_scores
```

Рис. 2. Приклад реалізації PageRank для визначення важливості вузлів на мові Python

Ще один важливий клас алгоритмів — це ті, що аналізують розподіл і сегментацію мережі. Ці алгоритми, включаючи каскадні моделі (рис. 3) та алгоритми поширення етикеток, допомагають зрозуміти, як інформація або поведінка поширюється серед користувачів. Вони корисні для моделювання процесів, таких як вірусний маркетинг або поширення соціальних інновацій, де важливо передбачити швидкість та шляхи розповсюдження інформаційних потоків.

```python
def cascade_model(G, seed_nodes, steps=3):
    active = set(seed_nodes)
    for _ in range(steps):
        new_active = set()
        for node in active:
            for neighbor in G.neighbors(node):
                new_active.add(neighbor)
        active.update(new_active)
    return active


active_nodes = cascade_model(G, ['A'])
print( "Активні вузли після каскадх:", active_nodes)
```

Рис. 3. Демонстрація використання алгоритму розподілу мережі на мові Python

Для дослідження було обрано датасет спільноти "Ukrainian Conflict" у СМ Reddit. Датасет було створено для аналізу соціальної активності та інформаційних тенденцій під час повномасштабного вторгнення в Україну – "News & Events Surrounding Russia's Invasion of Ukraine". Джерело даних – Kaggle (https://www.kaggle.com/datasets/gpreda/russian-invasion-of-ukraine).

На **першому етапі** вхідний набір повідомлень було приведено у структурований та однорідний вигляд, придатний для подальшого аналізу тексту, часової динаміки, кластеризації та визначення настроїв. У результаті очищення було сформовано табличну структуру з ключовими колонками: title (заголовок), text (об'єднаний текст), domain (домен джерела), timestamp (час публікації) та source_dataset (назва датасету).

На **другому етапі** здійснено *тематичний аналіз* текстового контенту, що дозволяє виявити ключові теми і напрями обговорень. Автоматичне групування контенту за змістом дозволяє отримати уявлення про домінуючі наративи. Для тематичного моделювання використано алгоритм Latent Dirichlet Allocation, один із найпопулярніших підходів до виділення тем у великих масивах тексту. Перед цим текстові дані були векторизовані за допомогою CountVectorizer з урахуванням англійських стоп-слів , а також згенерованих біграм і триграм (словосполучень із двох і трьох слів), які дозволяють виявити не лише окремі ключові слова, а й сталу лексику та важливі фрази. Наприклад, замість простого "war" модель може зафіксувати фрази на кшталт "russian invasion", "military aid" або "civilian casualties". Було обрано значення n_components=5, щоб модель могла виділити п'ять основних тем. Після навчання LDA-моделі зібрані теми були розшифровані шляхом аналізу топ-10 найвагоміших слів для кожної з них (рис. 4). Ці слова описують зміст кожної теми та ідентифікують її, наприклад: "воєнні дії", "міжнародна підтримка", "гуманітарна криза" тощо. У результаті тематичного аналізу були виявлені ключові інформаційні сюжети, які переважають у Reddit-спільноті. Це дозволяє зробити подальші висновки щодо того, які саме теми привертали найбільшу увагу користувачів у певні періоди часу, а також виявити потенційні точки емоційного або політичного напруження. Тематичне моделювання стало основою для подальших етапів, зокрема, оцінки емоційного тону повідомлень та динаміки інформаційних хвиль.

Топ-5 тем:

Topic	Top Words
1	comment, russia, like, china, good, gas, just, russian, yes, oil
2	https, com, comments, ukrainianconflict, www, https www, reddit, reddit com, www reddit, com ukrainianconflict
3	russia, ukraine, war, comment, russian, nato, putin, military, ukrainian, germany
4	comment, just, don, like, people, think, putin, know, russians, really
5	russian, ukraine, ukrainian, forces, comment, artillery, russia, military, kherson, missiles

Рис. 4. Топ-5 тем та ключові слова по кожній темі

На **третьому етапі** проведено аналіз емоційного тону повідомлень для розуміння настроїв, характеру обговорень та емоційного впливу ключових тем. Враховуючи великий обсяг даних, було обрано підмножину з 200000 випадково вибраних постів для демонстрації підходу. Аналіз здійснювався за допомогою бібліотеки TextBlob, яка на основі лінгвістичних правил обчислює полярність кожного повідомлення в діапазоні від –1 (негативне) до +1 (позитивне). Було створено функцію, яка класифікує кожне повідомлення як позитивне, негативне або нейтральне, залежно від рівня полярності. Повідомлення з показником полярності більше однієї десятої вважались позитивними, менше мінус однієї десятої — негативними, а всі інші — нейтральними. Цей підхід дозволив швидко отримати уявлення про загальний емоційний фон в обговореннях. Результати показали, що більшість повідомлень мали нейтральний тон — близько 56.6%, приблизно 29.5% повідомлень – як позитивні, а 13.9% – негативні. Це може свідчити про певну збалансованість у тональності дискусій, а також про наявність значного емоційного забарвлення — як у бік підтримки, так і в бік тривоги, гніву чи обурення. Для глибшого аналізу здійснено візуалізацію динаміки емоцій у часі. Повідомлення було згруповано за датами, і для кожного дня обчислювалась кількість позитивних, негативних та нейтральних постів. Побудований графік показав (рис. 5), що в більшості днів переважають нейтральні повідомлення, однак періодично з'являються чітко виражені піки емоційної активності. У деякі дні спостерігалось різке зростання негативного або позитивного тону, такі сплески можуть бути пов'язані з важливими подіями, наприклад, наступами, міжнародною допомогою або виявленням воєнних злочинів. Подальша деталізація цих дат за допомогою тематичного аналізу дозволяє точніше визначити, які саме події викликали відповідну реакцію.

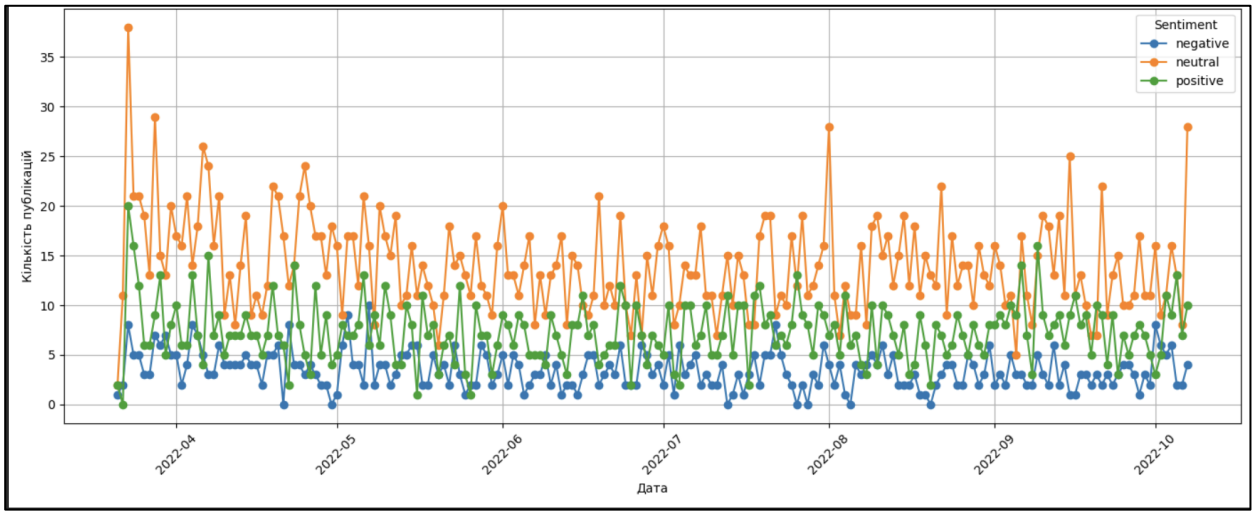

Рис. 5. Динаміка емоційного забарвлення повідомлень

На **четвертому етапі** було проаналізовано динаміку кількості публікацій в СМ (рис.6).

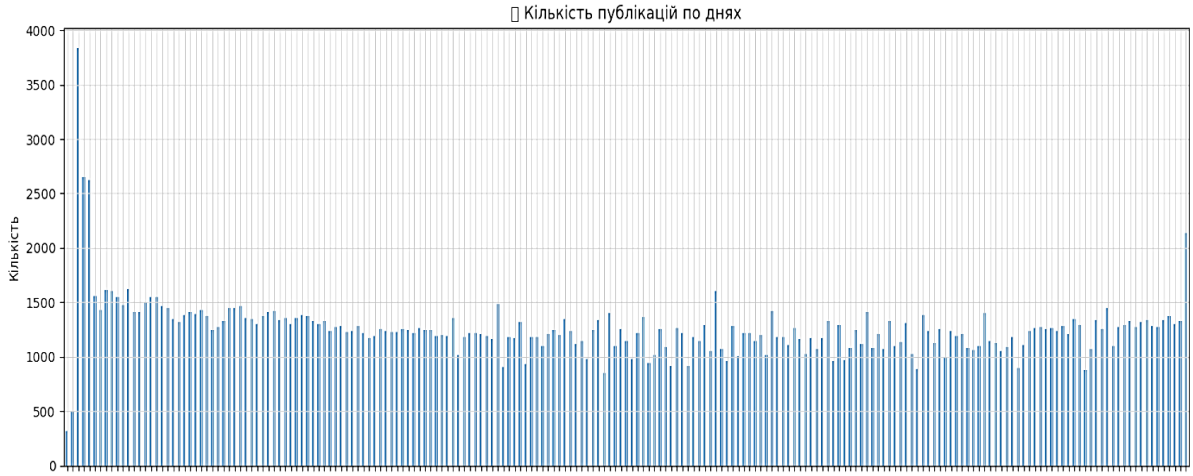

Рис. 6. Динаміка щоденної кількості публікацій

На основі отриманих даних було виявлено найпопулярніші теми у найактивніші дні.

На **п'ятому етапі** було більш детально проаналізовано зміст та ключові інформаційні тренди. Цей етап передбачав: з'ясувати, які слова найчастіше вживаються у позитивних та негативних повідомленнях, та відстежити, як змінювалася частота згадок важливих тем і персоналій у часі. Перший підетап включав створення "хмар слів" для двох груп повідомлень – позитивних і негативних. Хмари слів дозволяють візуально і швидко виявити найбільш уживані лексеми. У позитивному кластері домінували слова на кшталт support, help, thank, victory, freedom — що вказує на високий рівень уваги до тем міжнародної підтримки, волонтерства, опору та гуманітарної допомоги. Це також свідчить про те, що, попри загрозливу ситуацію, користувачі активно обговорювали приклади солідарності, позитивні зміни та спільну боротьбу. У негативних публікаціях переважали слова на зразок attack, war, killed, russia, missile. Така різниця у лексичних ядрах підтверджує ефективність аналізу настроїв та підкреслює емоційно-смислову полярність інформаційного контенту.

Рис. 7. Хмари слів розподілені по емоційному забарвленню

У другому підетапі було реалізовано тренд-постинг за ключовими словами – відстеження, як часто згадувалися певні слова в постах у різні дати. Для аналізу було обрано 8 ключових слів: Zelensky, putin, Bucha, missile, refugee, nato, russia, Crimea (рис.8).

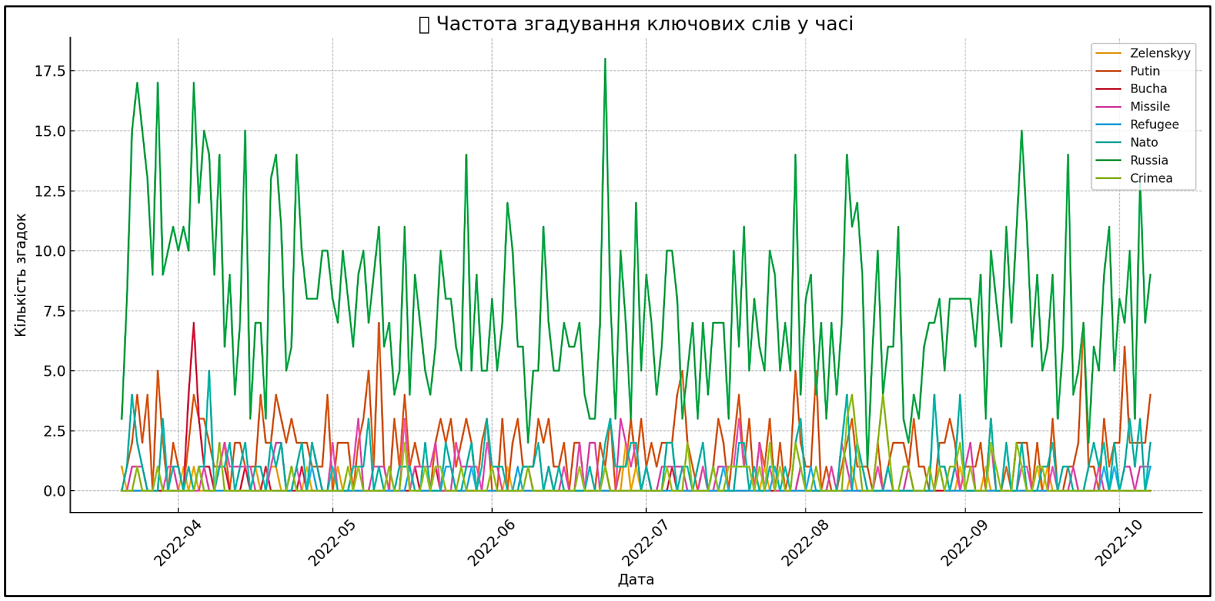

Рис. 8. Динаміка частоти згадування ключових слів у часі

Аналіз основних метрик. Побудуємо граф, який моделює зв'язки між доменами новинних сайтів та ключовими словами, отриманими з тематичного моделювання (LDA). У графі вузлами виступають домени та ключові слова, що найчастіше зустрічаються в текстах. Ребра між ними відображають факт згадки відповідного ключового слова в текстах новин відповідного домену. Вага ребра відповідає кількості таких згадок. Таким чином, побудована мережа ілюструє інформаційні зв'язки між тематиками та джерелами поширення. Після побудови графу проведемо його кількісний опис через основні метрики СМ. Також для можливості подальшого аналізу визначимо загальну кількість вузлів і ребер, щільність графу, яка дозволяє оцінити, наскільки тісно зв'язана мережа (рис. 9).

```
Кількість вузлів: 2518
Кількість ребер: 10896
Щільність мережі: 0.0034
Кількість компонент зв'язності: 171
```

Рис. 9. Основні характеристики мережі

Було виділено компоненти зв'язності, тобто групи вузлів, кожен з яких може бути досягнутий з будь-якого іншого у межах тієї ж групи. Це показує фрагментованість або інтеграцію інформаційного простору. Далі було обчислено центральності – ключові показники впливовості вузлів та експортовано у CSV-файл. Було визначено середній коефіцієнт кластеризації, який

показує, наскільки ймовірно, що вузли формують тісні групи. Висока кластеризація може свідчити про наявність інформаційних "клубів" чи сегментів зі спільними тематиками (рис. 10).

	Node	Degree Centrality	Closeness Centrality	Betweenness Centrality
1	russia	0.6348827969805324	0.5741033375981417	0.3228646941781116
2	ukraine	0.5343663090981327	0.5068276801584336	0.2764062784847208
3	russian	0.4513309495431069	0.46209500061742315	0.11085528518681079
4	war	0.354787445371474	0.4190888400850471	0.09627013337130134
5	ukrainian	0.26618990862137465	0.3861118424354474	0.07116064021990177
6	putin	0.2391736193881605	0.37706442503861404	0.05888375552314844
7	military	0.14978148589590784	0.34993315045156953	0.01531533473835
8	don	0.14342471195868098	0.3481517535673109	0.014854545311592504
9	russians	0.12514898688915374	0.3431298091759354	0.005481231084189923

Рис. 10. Обчислені центральності

Після цього було реалізовано пошук найкоротших шляхів між вузлами, що дозволяє моделювати, як швидко може передаватися інформація від одного джерела до іншого. Для виявлення структурованих груп у графі застосовано алгоритм Girvan-Newman, який дозволяє знайти спільноти вузлів – підмережі з більш щільними внутрішніми зв'язками. Це було зроблено для виявлення тематичних або джерельних "кластерів", які функціонують більш ізольовано.

Наступним етапом було застосування алгоритму PageRank для ранжування (див п. 4.2.4) веб-сторінок та визначення найвпливовіших вузлів у графі (рис. 11). Він оцінює значущість вузла з урахуванням кількості та важливості вузлів, що на нього посилаються. За допомогою цього визначаються ключові джерела інформації та теми з найбільшим охопленням.

	Node	PageRank
1	russia	0.06352698437420336
2	ukraine	0.04600124092763398
3	russian	0.033441488141522833
4	war	0.02337870590199336
5	ukrainian	0.018257129923538926
6	putin	0.013905960743999952
7	don	0.012777413146355603
8	like	0.0109784469411934
9	just	0.010536603807118301
10	military	0.006723637206836556
11	people	0.0061231993689521876
12	russians	0.006049356399757399
13	think	0.0053367540537537
14	know	0.005052967639156428
15	nato	0.004927765164273035

Рис. 11. Найвпливовіші вузли

Після цього було виявлено мости – такі ребра, видалення яких розділяє мережу на нові компоненти. Це використовується для того, щоб виявити основні зв'язки в інформаційній структурі, які є важливими для збереження цілісності мережі, і без яких мережа не буде функціонувати цілісно. Кількість мостів у мережі — 479.

```
 <-> reddit com
www.humblebundle.com <-> ukraine
goukraine.ytmnd.com <-> ukraine
www.sewerynszwarocki.pl <-> com
digital.library.unt.edu <-> know
www.lapresse.ca <-> just
www.ua-football.com <-> military
brcf-ua.blogspot.com <-> ukrainian
dronexl.co <-> ukraine
www.ar500armor.com <-> ukraine
en.media.renaultgroup.com <-> russia
kostroma.news <-> com
www.motorbiscuit.com <-> ukrainian
politicsheadlines.com <-> russia
ru.usembassy.gov <-> war
```

Рис. 12. Виявлені мости

Результати інтегровано у графову БД Neo4j, щоб мати можливість візуально оцінити граф та наявні зв'язки і виявити певні аномалії чи закономірності, які важко пояснити математично або за допомогою певних алгоритмів. Усі вузли і зв'язки були збережені разом з атрибутами, такими як тип вузла та вага зв'язку. Таким чином надалі можна проводити запити і аналіз безпосередньо у середовищі Neo4j, використовуючи Cypher-запити.

Створимо інтеграцію графової моделі networkx у середовище Neo4j, щоб зберігати, візуалізувати та аналізувати структури СМ через Cypher-запити. Спочатку створюємо вузли графу. Кожен вузол у networkx може мати тип (наприклад, "domain" або "keyword"). Якщо тип не вказано, використовується загальна мітка "Entity". Вузол створюється з відповідною міткою та іменем. Створимо зв'язки між вузлами. Кожне ребро перетворюється у зв'язок типу "MENTIONS", що означає, що певний домен згадує певне ключове слово. Додається також атрибут weight, який відображає кількість таких згадок, цей атрибут потрібен для зваженого аналізу мережі. Використовуватимемо Cypher-запити для дослідження, наприклад: пошук найбільш згадуваних ключових слів, побудова підмереж певного домену, або аналіз ваги зв'язків у певній тематиці.

Розглянемо приклади застосування Cypher-запитів для дослідження побудованого графу. Виконаємо пошук найбільш впливових доменів (вузлів з найбільшою кількістю зв'язків). Напишемо запит, який реалізовує концепцію ступеневої центральності: буде знайдено такі вузли-домени, які мають найбільшу кількість вихідних зв'язків до ключових слів. У СМ це еквівалент користувачів чи джерел, що найактивніше взаємодіють з іншими, тобто продукують контент, пов'язаний з багатьма темами. Запит проходить по всіх вузлах із міткою domain, рахує кількість зв'язків типу MENTIONS, які виходять з кожного домену до ключових слів, і повертає топ-10 таких доменів за кількістю згадок. Це ідентифікує найбільш інформаційно активні сайти. Для цього виконаємо наступний запит:

```
MATCH (d:domain)-[r:MENTIONS]->()
RETURN d.name AS Domain, COUNT(r) AS Mentions
ORDER BY Mentions DESC
LIMIT 10
```

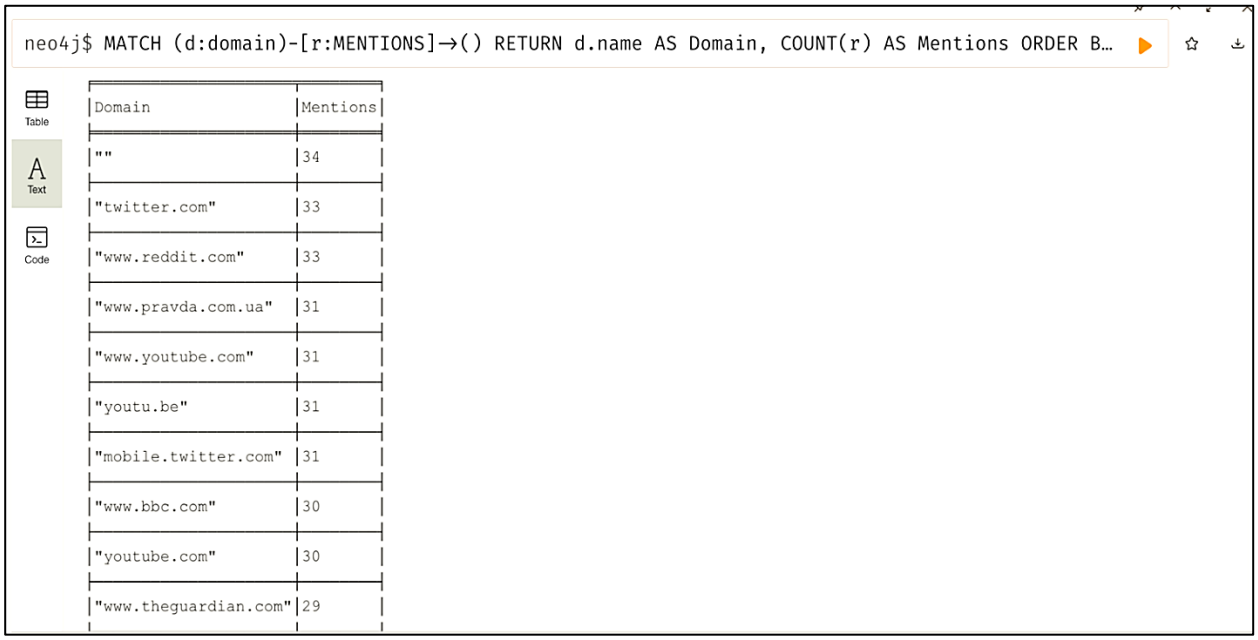

Рис. 13. Результат пошуку найбільш впливових доменів

Іншим прикладом є пошук ключових слів, які згадуються найбільшою кількістю доменів. Це допоможе виявити найбільш центральні концепції чи теми у мережі, тобто такі ключові слова, що об'єднують найбільшу кількість джерел. Це також наближається до поняття міжвузлової центральності, бо такі вузли можуть з'єднувати велику кількість доменів. Запит повертає ключові слова, що згадуються найбільшою кількістю доменів. Він проходить по всіх keyword вузлах і підраховує кількість вхідних зв'язків MENTIONS, що показує скільки джерел говорили про ту чи іншу тему. Запит виглядає наступним чином:

```
MATCH (k:keyword)<-[r:MENTIONS]-()
RETURN k.name AS Keyword, COUNT(r) AS MentionedByDomains
```

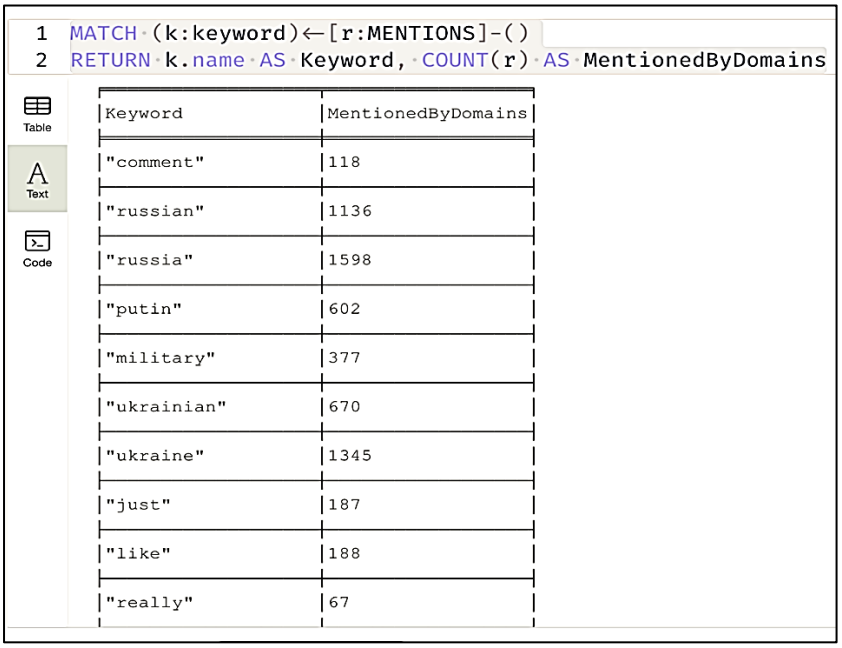

Рис. 14. Результат пошук ключових слів, які згадуються найбільшою кількістю доменів

Висновки і перспективи подальших розробок. Проведено дослідження методів аналізу СМ та проведено програмну реалізацію виявлення та оцінки характеристик, у т.ч. динамічних, окремих спільнот. Розроблений інструментарій дозволяє:

- виявляти потенційні ботоферми – групи акаунтів з високою щільністю внутрішніх зв'язків і майже відсутніми зовнішніми контактами, синхронізацією публікацій високої подібності;
- виявляти прихильників та опонентів певного наративу, ідеї, партії чи особи – кластери акаунтів, що поширюють позитивні/негативні повідомлення про об'єкт, порівняння середнього поляритету і топ ключових слів у кожному кластері, визначення ключових впливових вузлів у кожній групі;
- оцінювати структурну інтеграцію та згуртованість спільнот – розмір кластерів, коефіцієнт кластеризації, щільність зв'язків, виявлення мостів і точок зламу для оцінки стійкості інформаційного поширення.

Отримані результати підтверджують актуальність застосування графових БД, зокрема, Neo4j, у задачах аналізу складних соціальних структур.

Розроблене ПЗ можна адаптувати до інших тем, джерел або платформ та розширити його функціонал для різних типів соціальних мереж, продовжити дослідження у напрямах:
- автоматичне виявлення інформаційних хвиль та піків емоційної активності;
- розширення графових моделей шляхом включення часових зв'язків і взаємодій між користувачами;
- застосування методів машинного навчання до класифікації тем і виявлення бот-активності;
- адаптація системи для моніторингу інших СМ.

Результати дослідження можуть бути використані для моніторингу соціальних настроїв, виявлення інформаційних атак та центрів впливу у журналістиці та мас-медіа, дослідженнях громадської думки, розробці систем раннього попередження в інформаційній безпеці, маркетингу, PR тощо.

Список бібліографічного опису

1. The Global Risks Report 2024. 19th Edition. URL: https://www.weforum.org/publications/global-risks-report-2024/ (дата звернення: 11.08.2025).
2. Мазуренко В.В., Штовба С.Д. (2015) Огляд моделей аналізу соціальних мереж. *Вісник Вінницького політехнічного інституту*, No.2. С.62–74.
3. Singh, S., Samya, M., Srivastava D., Shakya H., Kumar N. (2024) Social Network Analysis: A Survey on Process, Tools, and Application. *ACM Comput. Surv*, Vol.56, #8, Article No192, p. 1 – 39, https://doi.org/10.1145/3648470
4. Мелешко Є. В. (2019) Методи кластеризації графів соціальних мереж для побудови рекомендаційних систем. *Системи управління, навігації та зв'язку*, No 54, Т.2. С.129-134. https://doi.org/10.26906/SUNZ.2019.2.129
5. Chunaev, P. (2020) Community detection in node-attributed social networks: A survey. *Computer Science Review*, Vol. 37, Art. No. 100286. https://doi.org/10.1016/j.cosrev.2020.100286
6. Shen, A., Kam-Pui, C. (2024). Community Detection Framework Using Deep Learning in Social Media Analysisю *Applied Sciences* 14, no. 24: 11745. https://doi.org/10.3390/app142411745
7. Nooribakhsh, M., Fernández-Diego, M., González-Ladrón-De-Guevara, F., Mollamotalebi, M. (2024) Community detection in social networks using machine learning: a systematic mapping study. *Knowledge and Information Systems*, Vol. 66, p.7205–7259. https://doi.org/10.1007/s10115-024-02201-8
8. Alotaibi, N., Rhouma, D. (2022) A review on community structures detection in time evolving social networks. *Journal of King Saud University - Computer and Information Sciences*, Vol. 34, Issue 8, Part B, p. 5646-5662. http://dx.doi.org/10.1016/j.jksuci.2021.08.016
9. Zhao, Y., Bai, W., Qiao, T., Wang, W. (2025) Modularity of online social networks acts as a reliable predictor of both whole-network and ego-network characteristics over time. *Humanities and Social Sciences Communications*, Vol. 12, Article number: 839. https://doi.org/10.1057/s41599-025-05022-4
10. Goglia, D., Vega, D. (2024) Structure and dynamics of growing networks of Reddit threads. *Applied Network Science*, Vol.9, 48. https://doi.org/10.1007/s41109-024-00654-y
11. Duzen, Z., Riveni, M., Aktas, M. (2023) Analyzing the Spread of Misinformation on Social Networks: A Process and Software Architecture for Detection and Analysis. *Computers*, Vol. 12(11), 232; https://doi.org/10.3390/computers12110232
12. Monteiro, J., Sá, F., & Bernardino, J. (2023). Experimental Evaluation of Graph Databases: JanusGraph, Nebula Graph, Neo4j, and TigerGraph. *Applied Sciences*, 13(9), 5770. https://doi.org/10.3390/app13095770

References

1. The Global Risks Report 2024. 19th Edition. URL: https://www.weforum.org/publications/global-risks-report-2024/ (Access date: 11/08/2025).
2. Mazurenko V., Shovba S. (2015) Overview of models for social network analysis. *Bulletin of Vinnytsia National Technical University*, No.2, p.62–74.
3. Singh, S., Samya, M., Srivastava D., Shakya H., Kumar N. (2024) Social Network Analysis: A Survey on Process, Tools, and Application. *ACM Comput. Surv*, Vol.56, #8, Article No192, p. 1 – 39, https://doi.org/10.1145/3648470
4. Meleshko Ye. (2019) Graph clustering methods in social networks for building recommendation systems. Control, Navigation and Communication Systems. Academic Journal, No 54, Part 2, p. 129-134. https://doi.org/10.26906/SUNZ.2019.2.129

5. Chunaev, P. (2020) Community detection in node-attributed social networks: A survey. *Computer Science Review*, Vol. 37, Art. No. 100286. https://doi.org/10.1016/j.cosrev.2020.100286
6. Shen, A., Kam-Pui, C. (2024). Community Detection Framework Using Deep Learning in Social Media Analysisю *Applied Sciences* 14, no. 24: 11745. https://doi.org/10.3390/app142411745
7. Nooribakhsh, M., Fernández-Diego, M., González-Ladrón-De-Guevara, F., Mollamotalebi, M. (2024) Community detection in social networks using machine learning: a systematic mapping study. *Knowledge and Information Systems*, Vol. 66, p.7205–7259. https://doi.org/10.1007/s10115-024-02201-8
8. Alotaibi, N., Rhouma, D. (2022) A review on community structures detection in time evolving social networks. *Journal of King Saud University - Computer and Information Sciences*, Vol. 34, Issue 8, Part B, p. 5646-5662. http://dx.doi.org/10.1016/j.jksuci.2021.08.016
9. Zhao, Y., Bai, W., Qiao, T., Wang, W. (2025) Modularity of online social networks acts as a reliable predictor of both whole-network and ego-network characteristics over time. *Humanities and Social Sciences Communications*, Vol. 12, Article number: 839. https://doi.org/10.1057/s41599-025-05022-4
10. Goglia, D., Vega, D. (2024) Structure and dynamics of growing networks of Reddit threads. *Applied Network Science*, Vol.9, 48. https://doi.org/10.1007/s41109-024-00654-y
11. Duzen, Z., Riveni, M., Aktas, M. (2023) Analyzing the Spread of Misinformation on Social Networks: A Process and Software Architecture for Detection and Analysis. *Computers*, Vol. 12(11), 232; https://doi.org/10.3390/computers12110232
12. Monteiro, J., Sá, F., & Bernardino, J. (2023). Experimental Evaluation of Graph Databases: JanusGraph, Nebula Graph, Neo4j, and TigerGraph. *Applied Sciences*, 13(9), 5770. https://doi.org/10.3390/app13095770