INFORMATION TECHNOLOGY

DOI 10.20535/2411-1031.2025.13.1.328753
UDC 004.9

ОЛЕКСАНДР ПУЧКОВ,
ДМИТРО ЛАНДЕ,
ІГОР СУБАЧ

МЕТОДИКА СТВОРЕННЯ, КЛАСТЕРИЗАЦІЇ ТА ВІЗУАЛІЗАЦІЇ КОРЕЛЯЦІЙНИХ МЕРЕЖ, ЩО ВИЗНАЧАЮТЬСЯ ДИНАМІКОЮ ТЕМАТИЧНИХ ІНФОРМАЦІЙНИХ ПОТОКІВ

В умовах стрімкого зростання обсягів інформації, яка циркулює в соціальних медіа та інтернет-просторі, виникає гостра потреба в ефективних методах аналізу та візуалізації тематичних інформаційних потоків. Кореляційні мережі є потужним інструментом для формалізації таких процесів, оскільки вони дозволяють виявляти взаємозв'язки між різними об'єктами, у тому числі, на основі аналізу їхньої динаміки. Особливо це актуально для сфери кібербезпеки, де оперативне виявлення тенденцій та зв'язків між подіями може мати вирішальне значення. Стаття присвячена розробці методики створення, кластеризації та візуалізації кореляційних мереж, які визначаються динамікою тематичних інформаційних потоків. Пропонується підхід, що базується на аналізі векторів динаміки публікацій, отриманих за допомогою систем контент-моніторингу соціальних медіа. Кореляційні мережі формуються на основі взаємозв'язків між векторами, що відображають розподіл документів за датами. Для візуалізації та аналізу мереж використовуються інструменти, такі як Gephi, а також запропонована авторська діаграма Ph-Di для відображення динаміки інформаційних потоків. Методика дозволяє виявляти групи взаємопов'язаних об'єктів, що може бути корисним для аналізу тематичних інформаційних потоків, зокрема в сфері кібербезпеки. Результати дослідження можуть слугувати основою для побудови ймовірнісних мереж та подальшого сценарного аналізу. Переваги запропонованої методики полягають у низькій розмірності векторів, що спрощує їх обробку та аналіз, незалежності від мови, завдяки чому методика може бути застосована для аналізу інформаційних потоків різними мовами, а також простоті реалізації, що робить її доступною для широкого кола дослідників та аналітиків у сфері кібербезпеки.

Ключові слова: кореляційні мережі, тематичні інформаційні потоки, кластеризація, візуалізація, вектори динаміки, контент-моніторинг, кібербезпека, Gephi, діаграма Ph-Di, семантичні мережі

Постановка проблеми. У сучасних умовах стрімкого зростання обсягів інформації, особливо в соціальних медіа та інтернет-просторі, виникає потреба в ефективних методах аналізу та візуалізації тематичних інформаційних потоків. Кореляційні мережі є потужним інструментом для формалізації таких процесів, оскільки вони дозволяють виявляти взаємозв'язки між різними об'єктами на основі аналізу їхньої динаміки [1]. Особливо це актуально для сфери кібербезпеки, де оперативне виявлення тенденцій та зв'язків між подіями може мати вирішальне значення.

Крім того, кореляційні мережі відіграють важливу роль у формалізації процесів аналізу тематичних інформаційних потоків. У таких мережах зв'язки між вузлами відображають значення кореляцій між векторами параметрів, що відповідають цим вузлам [2]. Для створення мережевих структур кожного вузла (тематики) формуються вектори параметрів – числові масиви, які відповідають тематичним документальним добіркам. Для цього застосовуються системи контент-моніторингу соціальних медіа.

Варто зауважити, що кореляція не завжди свідчить про причинно-наслідкові зв'язки, тому кореляційні мережі не слід розглядати як каузальні семантичні карти [3]. Проте кореляцію, поряд з іншими критеріями, можна використовувати як основу для ймовірнісних оцінок. Це означає, що кореляційні мережі можуть слугувати базою для створення ймовірнісних мереж і застосування технологій нечітких семантичних мереж для подальшого сценарного аналізу.

Відповідно до наведеного, доцільно розробити методику, яка зіставляє тематичному масиву документів, так званий, вектор динаміки, що відображає розподіл документів за датами. Кожному дню відповідає числове значення – кількість документів із тематичного масиву, причому розмірність цього вектору визначається кількістю днів, протягом яких формувався масив документів за заданою темою.

Під час створення мережевих структур для кожного об'єкта чи тематики формуються вектори, які їх представляють. Для цього передбачається використання системи моніторингу контенту в соціальних медіа, таких як, наприклад, "Кіберагрегатор" [4], [5] або InfoStream [6]. Подібні системи дають змогу отримувати масиви чисел, що відповідають тематичним добіркам документів. Ці масиви можна сформувати шляхом введення запиту через інтерфейс користувача з використанням інформаційно-пошукової мови.

Після побудови векторів, що відповідають окремим об'єктам, формується кореляційна мережа, яка використовується для збереження та візуалізації об'єктів, між якими існують об'єктивні зв'язки. Це дає змогу створювати вектори динаміки для різних об'єктів, взаємозв'язок між якими може бути неочевидним.

На рис. 1 представлено фрагмент вебінтерфейсу для отримання динаміки публікацій за тематикою "Кібербезпека". Запит до системи InfoStream [6] було сформульовано англійською мовою: Cybersecurity.

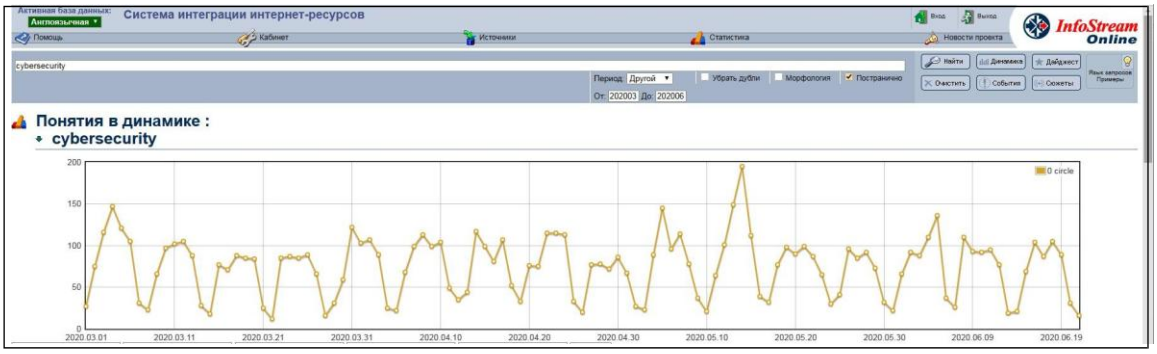

Рисунок 1 – Фрагмент інтерфейсу системи контент-моніторингу, де у вигляді графіка представлено вектор динаміки публікацій за тематикою "Кібербезпека"

Аналіз останніх досліджень і публікацій. З тематикою цієї статті пов'язані роботи різних авторів. Так в роботі [7] розглянуті питання використання можливостей системи Gephi для проведення дослідження, аналізу та візуалізації складних мереж.

У публікації [8] представлено всебічний емпіричний аналіз класифікації, ранжування та стабільності популярних алгоритмів на реальних наборах рейтингових даних за різних налаштувань.

Питанням дослідження методів кластеризації присвячено роботи [9], [10]. Кластеризація визначається як некероване навчання, в якому об'єкти групуються на основі певної подібності, притаманної їм. Розглянуто міри подібності, а також критерії оцінки, які є центральними компонентами методів кластеризаціїі. Зокрема, робота [10] присвячена практичному застосуванню одного з найпопулярніших і найпростіших алгоритмів кластеризації – K-середніх.

У роботі [11] запропоновано систему кореляційного мережевого аналізу для виявлення шахрайства, в якій використовується асиметрична кореляційна міра для розрізнення двох ролей, а саме, продавця-шахрая та покупця-шахрая.

У [12] описано методи виявлення спільнот (кластерів) у мережі та виявлення важливих або центральних об’єктів у мережевому графі.

Основним змістом роботи [13] є альтернативний спосіб визначення вузлів складної мережі з високою централізацією зв’язків, в якому вводиться нова метрика та рандомізований алгоритм для її оцінки. Емпірично показано, що вузли з високою центральністю, також, мають високу центральність міжвузлового зв’язку.

У публікаціях [14], [15] описано методи і технології дослідження, аналізу, та візуалізації графів і мереж.

У [16] наведено процес аналізу глобальної мережі дослідницьких спільнот за допомогою штучного інтелекту для отримання цінної інформації про їх структуру та вплив окремих робіт дослідників.

Формулювання цілей статті. Метою статті є розробка науково-методичного апарату для створення, кластеризації та візуалізації кореляційних мереж, які визначаються динамікою тематичних інформаційних потоків. Він має забезпечити ефективний аналіз взаємозв’язків між об’єктами, що представляють інтерес, та застосувати інформаційні технології для їх візуального представлення.

Для досягнення поставленої мети необхідно вирішити наступні взаємопов’язані часткові завдання:

1. Застосувати систему контент-моніторингу соціальних медіа для формування векторів динаміки публікацій, що відображають розподіл документів за датами.
2. Побудувати кореляційну мережу на основі взаємозв’язків між векторами динаміки.
3. Застосувати інструменти візуалізації для аналізу та представлення мережевих структур.
4. Розробити моделі для візуалізації динаміки інформаційних потоків у часовому розрізі.
5. Кластеризувати об’єкти на основі кореляційних зв’язків для виявлення груп, що є найбільш взаємопов’язаними.

Розроблений науково-методичний апарат може бути застосованим для аналізу різних тематичних потоків, зокрема, в сфері кібербезпеки, політики, соціальних явищ тощо, що робить його універсальним інструментом для дослідників та аналітиків.

Виклад основного матеріалу дослідження. Основні етапи побудови мережі взаємозв’язків об’єктів показано на рис. 2.

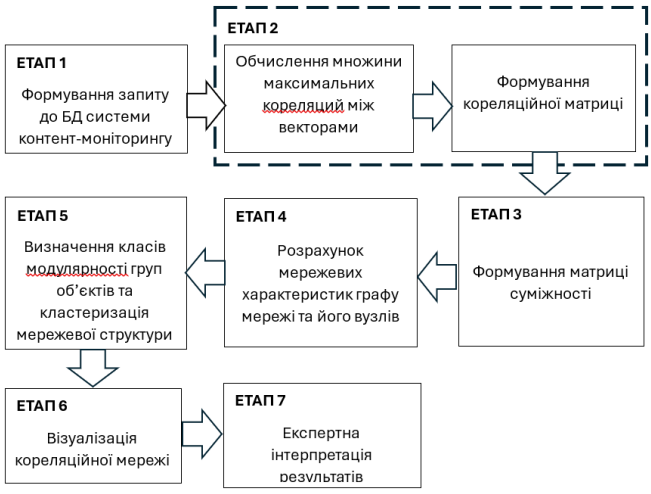

Рисунок 2 – Схема процесу формування, кластеризації та візуалізації кореляційних мереж, що визначаються динамікою тематичних інформаційних потоків

Етап 1. Для кожного об'єкту (поняття) формується запит інформаційно-пошуковою мовою системи контент-моніторингу до бази даних системи.

В якості прикладу розглянемо 7 об'єктів (тематик), запити про які наведені в таблиці 1. У системі, також, задається період пошуку, який визначає розмірність відповідних векторів динаміки.

Таблиця 1 – Приклади запитів про об'єкти кібербезпеки до системи InfoStream

№ об'єкту	Об'єкт	Запит	Знайдено документів (01.04.2024–10.06.2024)
1	Кібербезпека	Cyber~security	8979
2	Військові операції	Military~operation	2656
3	Інфляція	Inflation	10649
4	Коронавірус	Coronavirus|COVID-19	844498
5	Вибори в США	Elections&(USA|United~States)	21206
6	Протести в США	Protest&(USA|United~States)	20484
7	Терроризм	Terror	35631

У результаті опрацювання цих запитів визначається множина векторів динаміки, аналогічних тим, що наведені на рис. 1. Після цього виконується нормування цих векторів і, шляхом віконного згладжування елементів (з вікном спостереженні 7 діб), здійснюється позбавлення від тижневої періодичної складової.

Етап 2. Обчислюється множина максимальних кореляцій між векторами [1], що отримуються та формується кореляційна матриця з елементами в позначеннях формули (1):

$$a_{ij}(m) = \max_{m} \frac{\sum_{k=1}^{n-m} w_{k+m}^{i} w_{k}^{j}}{\sqrt{\sum_{k=m+1}^{n} \left(w_{k}^{i}\right)^{2}} \sqrt{\sum_{k=1}^{n-m} \left(w_{k}^{i}\right)^{2}}}, \quad (1)$$

де кожному об'єкту o_k, який належить множині $O = \{o_k\}_{k=1}^{|O|}$ відповідає вектор значень параметрів $\overline{w_k} = \left(w_1^k, w_2^k, \ldots, w_n^k\right)$, де n – кількість елементів у множині параметрів.

Функція max використовується з міркувань, що процеси, близькі за своєю суттю, мають схожу динаміку поведінки, хоча й можуть мати певний зсув.

Етап 3. Виконується формування матриці суміжності згідно з формулою (1) та збереження даних у файлі у форматі CSV. У таблиці суміжності, яка відображає зв'язки між усіма вузлами, відповідно до [1], ігноруються зв'язки, значення яких менші за визначений поріг. Вибір цього параметра залежить від досвіду аналітиків.

Після цього сформована матриця передається для обробки та візуалізації у систему аналізу графових структур, наприклад, систему Gephi [7], яка дозволяє виконувати дослідження даних, їх компоновку та візуалізацію. Варто зазначити, що матриця суміжності у форматі CSV для системи Gephi має певні особливості, які слід враховувати, а саме: вона містить нульові значення на діагоналі; значення розділяються символом ";" тощо (рис. 3).

Етап 4. Розраховуються мережеві характеристики графу. Система Gephi при обробці кореляційної мережі має ряд режимів, серед яких для отримання мережевих характеристик застосовується режим "Лабораторія даних". У цьому режимі можна розрахувати значення таких параметрів вузлів і графів, як PageRank, Hits, модулярність тощо. Відповідно, існують можливості ранжування вузлів матриці за цими параметрами (рис. 4).

	A	B	C	D	E	F	G	H
1		Cyber_Security	Military_Operation	Inflation	Coronavirus	Elections	Protest_USA	Terror
2	Cyber_Security	0.000	0.975	0.978	0.989	0.984	0.734	0.966
3	Military_Operation	0.975	0.000	0.979	0.980	0.975	0.750	0.965
4	Inflation	0.978	0.979	0.000	0.991	0.964	0.671	0.941
5	Coronavirus	0.989	0.980	0.991	0.000	0.981	0.701	0.950
6	Elections	0.984	0.975	0.964	0.981	0.000	0.763	0.972
7	Protest_USA	0.734	0.750	0.671	0.701	0.763	0.000	0.852
8	Terror	0.966	0.965	0.941	0.950	0.972	0.852	0.000

Рисунок 3 – Матриця суміжності наведеного прикладу в середовищі MS Excel

Id	Modularity Class
Cyber_Security	1
Military_Operation	0
Inflation	0
Coronavirus	1
Elections	2
Protest_USA	3
Terror	2

Рисунок 4 – Фрагмент таблиці даних в режимі “Лабораторія даних”

Етап 5. Здійснюється визначення класів модулярності груп об’єктів і подальша кластеризація завантаженої мережевої структури [15].

Модулярність мережі задається формулою (2):

$$Q = \frac{1}{2m}\sum_{v,w}\left[a_{vw} - \gamma\frac{k_v k_w}{2m}\right]\delta(c_v, c_w), \qquad (2)$$

де a_{ij} – елемент матриці суміжності A;

m – кількість ребер у графі;

k_v, k_w – степені вузлів v та w відповідно;

γ – розподільна здатність;

δ – дельта Кронекера (показує, чи знаходяться вузли c_v та c_w в одному модулі).

Етап 6. Візуалізація мережі в системі Gephi. Результати візуалізації мережі об’єктів (тематик), що відповідають заданим у таблиці запитам, наведені на рис. 5.

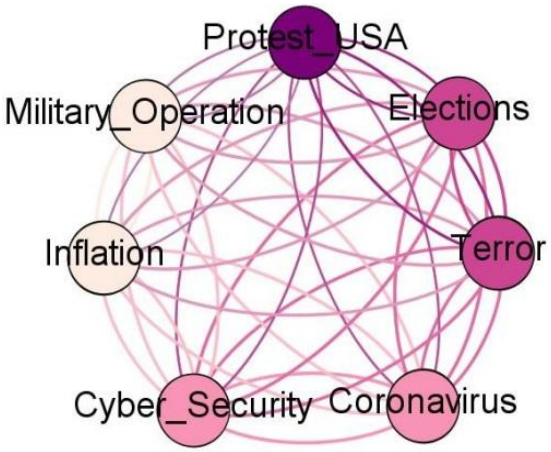

Рисунок 5 – Мережа сутностей (тематик) у середовищі системи Gephi

Етап 7. На останньому етапі здійснюється експертна інтерпретація результатів.

У свою чергу, для візуалізації динаміки появи об'єктів кібербезпеки в інформаційному просторі використовується набір чисел — кількість згадувань про ці об'єкти протягом p днів. На основі цього набору створюється спеціальна діаграма для відображення об'єктів у часі – Ph-Di (Phraseology Diagram) [16]. Ця діаграма має вигляд таблиці, де кожна клітинка зафарбована певним кольором, що відповідає частоті згадувань об'єкта i у день j. Рядки таблиці відповідають іменованим сутностям, які слугують своєрідними фільтрами інформаційного потоку, а стовпці – конкретним датам.

Таким чином, діаграма є двовимірною проекцією тривимірного набору часових рядів, що характеризують динаміку інформаційних потоків. При побудові діаграми Ph-Di можна здійснювати перестановку рядків (групування іменованих сутностей), що дозволяє виявляти важливі взаємозв'язки між об'єктами.

Для подальшого аналізу та кластеризації пропонується сформувати мережу зв'язків між іменованими сутностями на основі кореляції їх часових рядів. Це дозволяє виділяти групи найбільш взаємопов'язаних об'єктів та ідентифікувати, так звані, кліки (щільні підгрупи). Для цього використовується кореляційна матриця R розміру n×n, яка застосовується для виявлення найтісніших груп у просторі об'єктів кібербезпеки.

Отже, запропонована Ph-Di-діаграма дозволяє візуально аналізувати інформаційні потоки без необхідності додаткової обробки даних. Клітинки таблиці зафарбовані відповідно до обсягу публікацій про конкретний об'єкт у конкретний день: чим більше значення, тим світліший відтінок кольору. Це дає змогу легко ідентифікувати групи об'єктів, які найбільш пов'язані за інтенсивністю та часом публікацій про них (див. рис.6).

На рис. 6 наведено діаграму Ph-Di для об'єктів (концептів), що стосуються предметної області кібербезпеки. У даному випадку на діаграмі вертикальний вимір відповідає таким концептам, як суб'єкти кібербезпеки, а горизонтальний – датам публікацій про них. Колір комірок (точок) відповідає числовим значенням повідомлень за день щодо відповідних об'єктів кібербезпеки: світлі відтінки відповідають більшим значенням, темні – меншим. Горизонтальні світлі ризики на діаграмі відповідають періодам активності (згадуваності) відповідного об'єкта в соціальних мережах.

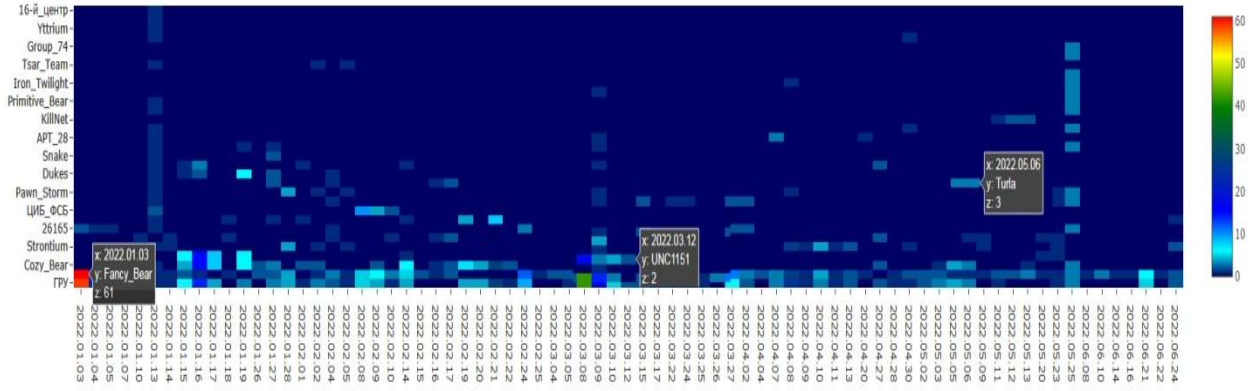

Рисунок 6 – Динаміка появи вибраних об'єктів кібербезпеки у часі

З формальної точки зору задача візуалізації динаміки появи об'єктів кібербезпеки в інформаційному просторі виглядає наступним чином.

Нехай маємо n об'єктів (сутностей), які представляють інтерес у контексті певної тематики з кібербезпеки. Для кожного об'єкта i ($i = 1,2,\ldots,n$) формується часовий ряд X_i, що відображає кількість згадувань цього об'єкта за p днів: $X_i = \left(x_{i1}, x_{i2}, \ldots, x_{ip}\right)$, де x_{ij} – кількість згадувань об'єкта i у день j.

Діаграма Ph-Di представляє собою двовимірну матрицю D розміром $n \times p$,

де: рядки матриці відповідають об'єктам ($i = 1,2,\ldots,n$);

стовпці матриці відповідають дням ($j = 1,2,\ldots, p$);

кожний елемент матриці d_{ij} відображає частоту згадувань об'єкта i у день j : $d_{ij} = x_{ij}$.

Відповідно до цього, колір кожної клітинки d_{ij} діаграми Ph-Di визначається значенням x_{ij} , де світліші відтінки відповідають більшим значенням, а темніші – меншим.

Для виявлення груп взаємопов'язаних об'єктів використовується кореляційна матриця R розміром $n \times n$, де кожен елемент r_{ik} визначає кореляцію між часовими рядами X_i та X_k: $r_{ik} = corr\left(X_i, X_k\right)$, де $corr\left(X_i, X_k\right)$ – коефіцієнт кореляції Пірсона між рядами X_i та X_k.

На основі кореляційної матриці R проводиться кластеризація об'єктів для виявлення груп (клік), які є найбільш взаємопов'язаними. Для цього можуть бути використані такі методи, як:

– ієрархічна кластеризація;
– k -середніх;
– виявлення спільнот у графах (community detection) тощо.

Результати кластеризації візуалізуються у діаграми Ph-Di так, що об'єкти, які належать до однієї групи, розташовуються поруч. Це дозволяє візуально ідентифікувати періоди активності та взаємозв'язки між ними.

Проілюструємо наведене на наступному прикладі. Нехай маємо 3 об'єкти ($n = 3$) та 5 днів ($p = 5$). Часові ряди для кожного об'єкта: $X_1 = \left(10,15,20,25,30\right)$, $X_2 = \left(5,10,15,20,25\right)$, $X_3 = \left(30,25,20,15,10\right)$.

Тоді двовимірна матриця D для діаграми Ph-Di має наступний вигляд:

$$D = \begin{pmatrix} 10 & 15 & 20 & 25 & 30 \\ 5 & 10 & 15 & 20 & 25 \\ 30 & 25 & 20 & 15 & 10 \end{pmatrix}.$$

Відповідно до неї, кореляційна матриця R буде наступною:

$$R = \begin{pmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 1 \end{pmatrix}.$$

На основі аналізу матриці R можна виявити, що об'єкти 1 та 2 мають високу позитивну кореляцію, тоді як об'єкт 3 має негативну кореляцію з ними.

Для практичної реалізації запропонованого науково-методичного апарату, діаграма Ph-Di створюється у форматі HTML-файлу з використанням мови JavaScript. Як видно з рисунка 6, яскраві горизонтальні лінії на діаграмі (що відповідають високій частоті окремих іменованих сутностей протягом певного періоду) вказують на тенденції та підвищену активність окремих об'єктів кібербезпеки в інформаційному полі Інтернету.

Вхідні дані для побудови діаграми Ph-Di подаються у вигляді CSV-файлу, де перший рядок містить дати, протягом яких спостерігались об'єкти кібербезпеки, а другий та наступні рядки відповідають конкретним об'єктам і включають загальну частоту, назву об'єкта та частоти згадувань цих об'єктів по днях (див. рис. 7).

Програма формування діаграми Ph-Di застосовує графічну бібліотеку з відкритим кодом plotly.js мови програмування JavaScript. Для зручності користувача, у вихідній формі реалізовано індикацію назви об'єкту і дати при інтерактивному наведенні маніпулятора миші на відповідну комірку діаграми Ph-Di.

Розроблений науково-методичний апарат, реалізований у вигляді методики, може бути використаний для аналізу різних тематичних потоків, зокрема в сфері кібербезпеки, політики, соціальних явищ тощо. Він дозволяє виявляти тенденції та взаємозв'язки між об'єктами, що є важливим для прийняття обґрунтованих рішень відповідними посадовими особами.

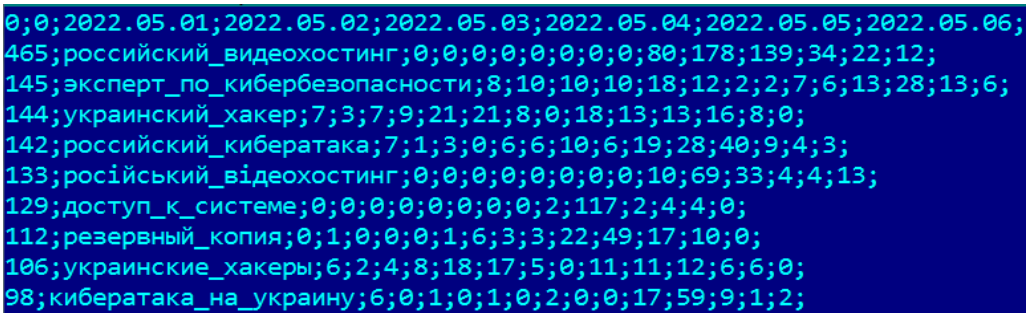

Рисунок 7 – Фрагмент вхідного CSV-файлу для побудови діаграми Ph-Di

Висновки. Таким чином, у рамках проведеного дослідження була розроблена методика створення, кластеризації та візуалізації кореляційних мереж, які визначаються динамікою тематичних інформаційних потоків. Дана методика демонструє ефективність під час аналізу взаємозв'язків між об'єктами, що представляють інтерес, та застосовує інструменти для їх візуального представлення.

У процесі дослідження було отримано низку ключових результатів, які підтверджують ефективність запропонованої методики та її практичну застосовність. Основні результати полягають у формалізації процесу аналізу інформаційних потоків, побудові кореляційних мереж, візуалізації та аналізу мереж, розробці діаграм Ph-Di та кластеризації об'єктів.

Розроблений науково-методичний апарат дозволяє систематизувати аналіз тематичних інформаційних потоків шляхом формування векторів динаміки публікацій. Ці вектори, що відображають розподіл документів за датами, є основою для подальшого аналізу та побудови кореляційних мереж.

На основі взаємозв'язків між векторами динаміки було створено кореляційні мережі, які дозволяють виявляти об'єкти, що об'єктивно пов'язані між собою. Це дає можливість аналізувати складні взаємозв'язки, які не завжди очевидні при традиційних підходах.

Використання інструментів візуалізації, таких як Gephi, дозволило ефективно аналізувати мережеві структури, розраховувати їх характеристики (наприклад, PageRank, модулярність тощо) та виявляти ключові вузли. Це спрощує інтерпретацію результатів та прийняття рішень на основі отриманих даних.

Діаграма Ph-Di стала потужним інструментом для візуалізації динаміки інформаційних потоків у часовому розрізі. Вона дозволяє візуально виявляти періоди активності об'єктів та групи взаємопов'язаних сутностей, що є особливо корисним для аналітиків.

Запропонована методика дозволила кластеризувати об'єкти на основі кореляційних зв'язків, що дає можливість виявляти групи, які найбільш взаємопов'язані. Це може бути використано для подальшого сценарного аналізу та прогнозування розвитку подій.

Таким чином, розроблена методика є ефективним інструментом для аналізу тематичних інформаційних потоків та може знайти широке застосування в різних галузях, зокрема в сфері кібербезпеки.

Переваги запропонованої методики полягають у низькій розмірності векторів, що спрощує їх обробку та аналіз, незалежності від мови, завдяки чому методика може бути застосована для аналізу інформаційних потоків різними мовами, оскільки вектори параметрів формуються на основі запитів до систем моніторингу, простоті реалізації, яка полягає у використанні готових програмних інструментів, таких як Gephi, Matlab або R, що робить методику доступною для широкого кола дослідників та аналітиків.

Перспективи подальших досліджень. Методика може бути використана аналітиками, дослідниками та фахівцями з кібербезпеки для моніторингу інформаційних потоків, виявлення тенденцій та прийняття обґрунтованих рішень. Вона також може бути інтегрованою у системи підтримки прийняття рішень для автоматизації процесів аналізу та візуалізації даних.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Д. Ланде, Л. Страшной, та І. Балагура, "Метод формування та кластеризації кореляційних мереж понять", *Реєстрація, зберігання і обробка даних*, № 23 (2), с. 27-36, 2012, doi: https://doi.org/10.35681/1560-9189.2021.23.2.239209.

[2] A.A. Snarskii, D.V. Lande, D.I. Zorinets, and A.V. Levchenko, "Reciprocally time correlating objects ranking", на *XVII Міжн. наук. конф. ім. Т.А. Таран інтел. ан. інф. (IAI 2017)*, Київ, 2017, с. 216-221.

[3] Д.В. Ланде, "Формування семантичної мапи понять в галузі парламентського контролю", *Інформація і право*, вип. 4 (47), с. 116-123, 2023, doi: https://doi.org/10.37750/2616-6798.2023.4(47).291611.

[4] D. Lande, O. Puchkov, and I. Subach, "Система аналізу великих обсягів даних з питань кібербезпеки із соціальних медіа", *Information Technologies and Security*, т. 8 (1), с. 4-18, 2020, https://doi.org/10.20535/2411-1031.2020.8.1.217993.

[5] D. Lande, O. Puchkov, and I. Subach, "Aggregation of information from diverse networks as the basis for training cyber security specialists on processing ultra-large data sets", *Information Technologies and Security*, т. 9 (1), с. 4-16, 2021, doi: https://doi.org/10.20535/2411-1031.2021.9.1.247256.

[6] Д.В. Ланде, *OSINT у кібербезпеці: навчальний посібник*. Київ, Україна: ТОВ "Інжиніринг", 2024.

[7] K. Cherven, *Mastering Gephi Network Visualization*. Birmingham, UK: Packt Publishing, 2015.

[8] G. Adomavicius, and J. Zhang, "Classification, ranking, and top-K stability of recommendation algorithms", *INFORMS Journal on Computing*, iss. 28 (1), pp. 129-147, 2016, doi: https://doi.org/10.1287/ijoc.2015.0662.

[9] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O.P. Patel, A. Tiwari, and C.-T. Lin, "A review of clustering techniques and developments", *Neurocomputing*, iss. 267, pp. 664-681, 2017, doi: https://doi.org/10.1016/j.neucom.2017.06.053.

[10] A.K. Jain, "Data clustering: 50 years beyond K-means", *Pattern Recognition Letters*, vol. 31, iss. 8, pp.651-666, 2010, doi: https://doi.org/10.1016/j.patrec.2009.09.011.

[11] P. Luo, K. Shu, J. Wu, L. Wan, and Y. Tan, "Exploring correlation network for cheating detection", *ACM Transactions on Intelligent Systems and Technology*, vol. 11, iss. 1, pp. 1-23, 2020, doi: https://doi.org/10.1145/3364221.

[12] A. Kassambara, *Network Analysis and Manipulation Using R: Quick Start Guide*. STHDA, 2017.

[13] N. Masuda, Z.M. Boyd, D. Garlaschelli, and P.J. Mucha, "Correlation networks: Interdisciplinary approaches beyond thresholding", *arXiv preprint arXiv:2311.09536*, 2023.

[14] T. Triantoro, "Graph Viz: Exploring, analyzing, and visualizing graphs and networks with Gephi and ChatGPT", in *ODSC Community*, March 30, 2023. [Online]. Available:https://opendatascience.com/graph-viz-exploring-analyzing-and-visualizing-graphs-and-networks-with-gephi-and-chatgpt. Accessed on: Feb. 19, 2025.

[15] Y. Liu et al., "Revisiting modularity maximization for graph clustering: A contrastive learning perspective", in *Proc. 30th ACM SIGKDD Conf. on Know. Disc. and DM*, Barcelona, 2024, pp. 1968-1979, doi: https://doi.org/10.1145/3637528.3671967.

[16] M. Zgurovsky, D. Lande, K. Yefremov, O. Dmytrenko, A. Boldak, and A. Soboliev, "Extracting and identifying relationships of key phrases in information flows", in *Proc. 2022 IEEE 3rd Intern. Conf. on Sys. An. & Intell. Com. (SAIC)*, Kyiv, 2022, pp. 4-7, doi: https://doi.org/10.1109/SAIC57818.2022.9923019.

Стаття надійшла до редакції 12.02.2025.

REFERENCE

[1] D. Lande, L. Strashnoi, and I. Balagura, "Method for the formation and clustering of correlation networks of concepts", *Registration, Storage and Processing of Data*, vol. 23, iss. 2, pp. 27-36, 2012, doi: https://doi.org/10.35681/1560-9189.2021.23.2.239209.

[2] A.A. Snarskii, D.V. Lande, D.I. Zorinets, and A.V. Levchenko, "Reciprocally time correlating objects ranking", in *Proc. XVII Inter. Scien. Conf. Named After T.A. Taran "Intell. An. of Inform. (IAI 2017)"*, Kyiv, 2017, pp. 216-221.

[3] D. Lande, "Formation of a semantic map of concepts in the field of parliamentary control", *Information and Law*, iss. 4 (47), pp. 116-123, 2023, doi: https://doi.org/10.37750/2616-6798.2023.4(47).291611.

[4] D. Lande, O. Puchkov, and I. Subach, "A system for analyzing large amounts of data on cybersecurity from social media", *Information Technologies and Security*, vol. 8, iss. 1, pp. 4-18, 2020, https://doi.org/10.20535/2411-1031.2020.8.1.217993.

[5] D. Lande, O. Puchkov, and I. Subach, "Aggregation of information from diverse networks as the basis for training cyber security specialists on processing ultra-large data sets", *Information Technologies and Security*, vol. 9, iss. 1, pp. 4-16, 2021, doi: https://doi.org/10.20535/2411-1031.2021.9.1.247256.

[6] D. Lande, *OSINT in cybersecurity: A textbook.* Kyiv, Ukraine: "Engineering LTD", 2024.

[7] K. Cherven, *Mastering Gephi Network Visualization*. Birmingham, UK: Packt Publishing, 2015.

[8] G. Adomavicius, and J. Zhang, "Classification, ranking, and top-K stability of recommendation algorithms", *INFORMS Journal on Computing,* iss. 28 (1), pp. 129-147, 2016, doi: https://doi.org/10.1287/ijoc.2015.0662.

[9] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O.P. Patel, A. Tiwari, and C.-T. Lin, "A review of clustering techniques and developments", *Neurocomputing*, iss. 267, pp. 664-681, 2017, doi: https://doi.org/10.1016/j.neucom.2017.06.053.

[10] A.K. Jain, "Data clustering: 50 years beyond K-means", *Pattern Recognition Letters*, vol. 31, iss. 8, pp.651-666, 2010, doi: https://doi.org/10.1016/j.patrec.2009.09.011.

[11] P. Luo, K. Shu, J. Wu, L. Wan, and Y. Tan, "Exploring correlation network for cheating detection", *ACM Transactions on Intelligent Systems and Technology*, vol. 11, iss. 1, pp. 1-23, 2020, doi: https://doi.org/10.1145/3364221.

[12] A. Kassambara, *Network Analysis and Manipulation Using R: Quick Start Guide*. STHDA, 2017.

[13] N. Masuda, Z.M. Boyd, D. Garlaschelli, and P.J. Mucha, "Correlation networks: Interdisciplinary approaches beyond thresholding", *arXiv preprint arXiv:2311.09536*, 2023.

[14] T. Triantoro, "Graph Viz: Exploring, analyzing, and visualizing graphs and networks with Gephi and ChatGPT", in *ODSC Community*, March 30, 2023. [Online]. Available:https://opendatascience.com/graph-viz-exploring-analyzing-and-visualizing-graphs-and-networks-with-gephi-and-chatgpt. Accessed on: Feb. 19, 2025.

[15] Y. Liu et al., "Revisiting modularity maximization for graph clustering: A contrastive learning perspective", in *Proc. 30th ACM SIGKDD Conf. on Know. Disc. and DM*, Barcelona, 2024, pp. 1968-1979, doi: https://doi.org/10.1145/3637528.3671967.

[16] M. Zgurovsky, D. Lande, K. Yefremov, O. Dmytrenko, A. Boldak, and A. Soboliev, "Extracting and identifying relationships of key phrases in information flows", in *Proc. 2022 IEEE 3rd Intern. Conf. on Sys. An. & Intell. Com. (SAIC)*, Kyiv, 2022, pp. 4-7, doi: https://doi.org/10.1109/SAIC57818.2022.9923019.

OLEXANDR PUCHKOV,
DMYTRO LANDE,
IHOR SUBACH

A METHODOLOGY FOR CREATING, CLUSTERING AND VISUALIZING CORRELATION NETWORKS DETERMINED BY THE DYNAMICS OF THEMATIC INFORMATION FLOWS

Given the rapid growth of information circulating in social media and the Internet space, there is an urgent need for effective methods of analyzing and visualizing thematic information flows. Correlation networks are a powerful tool for formalizing such processes, as they allow identifying relationships between different objects, including by analyzing their dynamics. This is especially relevant for the cybersecurity sector, where prompt detection of trends and connections between events can be crucial. The article is devoted to the development of a methodology for creating, clustering and visualizing correlation networks determined by the dynamics of thematic information flows. The article proposes an approach based on the analysis of vectors of publication dynamics obtained through social media content monitoring systems. Correlation networks are formed based on relationships between vectors reflecting the distribution of documents by dates. To visualize and analyze the networks, tools such as Gephi are used, as well as the author's own Ph-Di diagram to display the dynamics of information flows. The methodology allows identifying groups of interconnected objects, which can be useful for analyzing thematic information flows, particular in the field of cybersecurity. The results of the study can serve as a basis for building probabilistic networks and further scenario analysis. The advantages of the proposed methodology are the low dimensionality of the vectors, which simplifies their processing and analysis, language independence, so that the methodology can be used to analyze information flows in different languages, and ease of implementation, which makes it accessible to a wide range of researchers and analysts in the field of cybersecurity.

Keywords: correlation networks, thematic information flows, clustering, visualization, dynamics vectors, content monitoring, cybersecurity, Gephi, Ph-Di diagram, semantic networks.

Пучков Олександр Олександрович, кандидат філософських наук, професор, начальник, Інститутут спеціального зв'язку та захисту інформації Національного технічного університету України "Київський політехнічний інститут імені Ігоря Сікорського", Київ, Україна, ORCID 0000-0002-8585-1044, iszzi@iszzi.kpi.ua.

Ланде Дмитро Володимирович, доктор технічних наук, професор, завідувач кафедри інформаційної безпеки, Навчально-науковий фізико-технічний інститут Національного технічного університету України "Київський політехнічний інститут імені Ігоря Сікорського", Київ,Україна, ORCID 0000-0003-3945-1178, dwlande@gmail.com.

Субач Ігор Юрійович, доктор технічних наук, професор, завідувач кафедри кібербезпеки і застосування інформаційних систем і технологій, Інститут спеціального зв'язку та захисту інформації Національного технічного університету України "Київський політехнічний інститут імені Ігоря Сікорського", Київ, Україна, ORCID 0000-0002-9344-713X, igor_subach@ukr.net.

Puchkov Oleksandr, PhD in philosophy, professor, head of the Institute of special communication and information protection at the National technical university of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine.

Lande Dmytro, doctor of technical sciences, professor, chair of the academic department of the information security, Educational and scientific physico-technical institute at the National technical university of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine.

Subach Ihor, doctor of technical sciences, professor, chair of the academic department of the cyber security and application of information systems and technologies, Institute of special communication and information protection at the National technical university of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine.