Науковий вісник НЛТУ України
Scientific Bulletin of UNFU
ISSN 1994-7836 (print)
ISSN 2519-2477 (online)
https://doi.org/10.36930/40360115
Article received: 16.12.2025, accepted: 10.02.2026
Article published: 26.02.2026
UDC 004.[056+912]:519.17
Corresponding author: M. I. Stasiuk mykola.stasiuk@lnu.edu.ua

Б. М. Павлишенко, М. І. Стасюк
Львівський національний університет ім. Івана Франка, Львів, Україна

АНАЛІЗ ТЕМАТИЧНИХ ЗВ'ЯЗКІВ У БАГАТОМОВНОМУ ГРАФІ ЗНАНЬ КІБЕРБЕЗПЕКИ

Розширено методику аналізу кібербезпекових текстових корпусів шляхом формалізації міжтематичних залежностей як окремого структурного рівня над двомовним графом знань, що дає змогу перейти від аналізу окремих сутностей до мережевого трактування тематичних взаємодій. На основі 200 україномовних та 200 англомовних документів сформовано двомовний граф знань, виконано семантичну кластеризацію сутностей і побудовано асоціативні правила за заданими пороговими значеннями *support*, *confidence* та *lift*. Отримані правила відображено на рівень тематичних кластерів для побудови орієнтованого зваженого графа тематичних зв'язків, де вагою ребра виступає *lift*. Показано, що збереження орієнтованості графа є важливим, оскільки ігнорування напряму призводить до втрати інформації про конфігурацію тематичних потоків. Унаслідок цього орієнтований зважений граф забезпечує відображення міжтематичних впливів. Запропоновано інтегрований показник важливості *ITS*, який поєднує центральність за посередництвом, власну векторну центральність та максимальне значення *lift* для кожного кластера, проведено аналіз його чутливості за різних конфігурацій вагових коефіцієнтів. Встановлено стабільність ранжування домінантних кластерів за всіх протестованих параметрів, що свідчить про узгодженість топологічних і статистичних характеристик. Зокрема, кластери, пов'язані зі захистом маршрутизаторів, шкідливим програмним забезпеченням та офісним програмним забезпеченням, одночасно виконують роль структурних "містків" і беруть участь у найсильніших асоціативних залежностях. Окрім цього, використання двох мов дозволило сформувати дещо щільнішу та ієрархічно організовану структуру графа тематичних зв'язків порівняно з україномовною моделлю, навіть попри те, що англомовна підмножина окремо не генерувала достатньо сильних правил. Отже, інтеграція україномовного та англомовного корпусів у процес побудови графа дозволила підсилити приховані міжкластерні залежності, що не досягали порогових значень у межах одномовного корпусу та дала змогу виявити структурний центр графа. Переваги результатів дослідження полягають у формалізації моделі міжкластерних залежностей і розробленні кількісного механізму ранжування, що можна використати для аналітичної інтерпретації та виявлення структурно домінантних тематичних напрямів у сфері кібербезпеки. Подальші етапи дослідження передбачають масштабування корпусу, аналіз динаміки тематичних потоків у часі та перевірку стійкості інтегрованого показника важливості на великих мережах.

Ключові слова: асоціативні правила; семантична кластеризація; міжмовне узгодження; мережеві метрики; структуризація інформації.

Вступ / Introduction

Для сфери кібербезпеки характерне постійне зростання обсягів текстових даних, що містять опис кіберінцидентів, вразливостей, шкідливого програмного забезпечення та дій суб'єктів атак. Значну частину цієї інформації подано у вигляді неструктурованих аналітичних звітів, технічних повідомлень і рекомендацій щодо усунення вразливостей. Це ускладнює автоматизовану інтеграцію таких даних і системний їх аналіз. У світовій науковій літературі активно розвиваються підходи до автоматизованого вилучення знань із таких текстів із використанням графів знань, великих мовних моделей LLMs (англ. *Large Language Models*) та методів багатомовного семантичного подання. Графи знань застосовують для структуризації інформації про загрози, вразливості, шкідливе програмне забезпечення та суб'єкти атак, забезпечуючи формалізоване подання зв'язків між ними. Великі мовні моделі продемонстрували високу ефективність у завданнях вилучення сутностей і відношень, а багатомовні трансформерні моделі – у формуванні спільного векторного простору для текстів різними мовами.

Водночас, більшість наявних досліджень зосереджена або на побудові одномовних графів знань, або на міжмовному узгодженні окремих сутностей. У низці робіт після побудови графа знань застосовують кластеризацію даних для виявлення тематичних груп сутностей. Проте аналіз структурних залежностей між такими кластерами та оцінювання їхнього значення в межах побудованого графа знань залишаються недостатньо розробленими. Це зумовлено тим, що більшість наявних рішень зосереджені або на рівні окремих сутностей і відношень, або на їх міжмовному зіставленні, не переходячи до формалізації зв'язків між тематичними групами та їх кількісного аналізу з використанням мереже-

Павлишенко Богдан Михайлович, д-р техн. наук, доцент, кафедра системного проектування.
Email: bohdan.pavlyshenko@lnu.edu.ua; https://orcid.org/0000-0001-9515-3488
Стасюк Микола Іванович, аспірант, кафедра системного проектування.
Email: mykola.stasiuk@lnu.edu.ua; https://orcid.org/0009-0006-5256-2103
Цитування за ДСТУ: Павлишенко Б. М., Стасюк М. І. Аналіз тематичних зв'язків у багатомовному графі знань кібербезпеки. Науковий вісник НЛТУ України. 2026, т. 36, № 1. С. 136–143.
Citation APA: Pavlyshenko, B. M., & Stasiuk, M. I. (2026). Analysis of thematic relationships in a multilingual cybersecurity knowledge graph. *Scientific Bulletin of UNFU*, 36(1), 136–143. https://doi.org/10.36930/40360115

вих і статистичних показників.

Актуальність дослідження зумовлена потребою у розробленні інструментарію, який дає змогу не тільки структурувати багатомовні текстові масиви у сфері кібербезпеки, а й виявляти міжтематичні залежності та оцінювати потенційну важливість окремих тематичних напрямів у межах цілісної мережевої моделі.

Об'єкт дослідження – семантичний аналіз тематичних зв'язків внаслідок структуризації багатомовних текстових корпусів у сфері кібербезпеки.

Предмет дослідження – методи і засоби побудови двомовних графів знань, які дадуть змогу здійснити формалізацію графа тематичних зв'язків і кількісно оцінити структурну важливість тематичних кластерів.

Мета роботи – розробити та формалізувати граф тематичних зв'язків як окремий рівень подання знань над двомовним графом знань кібербезпеки, що дає змогу здійснювати аналіз тематичних зв'язків і кількісно оцінювати структурну важливість тематичних кластерів.

Для досягнення зазначеної мети визначено такі основні *завдання дослідження*:

- формалізувати граф тематичних зв'язків на основі тематичних кластерів і асоціативних правил, що дасть змогу перейти від аналізу локальних зв'язків між сутностями до міжкластерного рівня представлення залежностей;
- визначити вагову структуру графа тематичних зв'язків, що дасть змогу кількісно відобразити силу та напрям міжтематичних залежностей;
- застосувати мережеві метрики для аналізу структурного значення тематичних кластерів, що забезпечить оцінку їхньої ролі у графі з урахуванням топологічних характеристик;
- розробити інтегрований показник оцінювання важливості тематичних кластерів, який допоможе узагальнити мережеві та статистичні характеристики у межах єдиної кількісної оцінки;
- провести порівняльний аналіз україномовного та двомовного графів тематичних зв'язків, які уможливлять оцінювання впливу міжмовного розширення корпусу на формування міжкластерних зв'язків та структурну роль тематичних кластерів.

Аналіз останніх досліджень та публікацій. Розроблення багатомовного графа знань MKG (англ. *Multilingual Knowledge Graph*) у сфері кібербезпеки – це складне, але критично важливе завдання для автоматизації процесу виявлення загроз та обміну даними між різними абонентами, що поєднує семантичні технології, машинне навчання та лінгвістику. Основна проблема полягає в тому, що термінологія (наприклад, "vulnerability", "вразливість", "schwachstelle") має бути пов'язана не тільки як переклад, а як єдина сутність із набором атрибутів.

У роботі [8] запропоновано підхід LLM-TIKG для автоматизованого вилучення сутностей і відношень із відкритих джерел кібербезпеки з подальшим формуванням графа розвідки загроз (англ. *Threat Intelligence Knowledge Graph*). Автори показали, що використання LLM дає змогу підвищити повноту та точність вилучення знань порівняно з традиційними методами.

У дослідженні [4] розглянуто підхід до побудови графа знань кібербезпеки з використанням навчання великих мовних моделей у контексті. Автори продемонстрували можливість формування якісного графа знань навіть за обмежених анотованих даних.

У роботі [9] автори дослідили інтеграцію LLM і графів знань для отримання розвідки кіберзагроз, на основі якої можна діяти. У роботі показано, що поєднання мовних моделей і графових структур дає змогу автоматизувати процес оброблення великих масивів текстових звітів.

Результатом роботи [6] стала запропонована модель CTIKG для побудови графа знань на основі статей про кібербезпеку з використанням LLM без попередньої ручної анотації. Автори продемонстрували ефективність автоматичного формування графа знань у реальних сценаріях.

У роботі [3] автор у огляді семантичних веб-технологій і графів знань у дослідженнях кіберзагроз вказують на важливість інтеграції гетерогенних даних у графове подання. Автор зазначає, що графи знань забезпечують основу для аналітики загроз, однак не розглядає методи структурного оцінювання тематичних кластерів.

Автор роботи [10] досліджує інтегровані підходи до досліджень кіберзагроз із використанням графів знань і методів штучного інтелекту. У роботі показано, що поєднання моделей оброблення природної мови і графових структур покращує автоматизоване вилучення знань.

Отже, у наведених вище роботах достатньо детально опрацьовано питання автоматизованого вилучення знань із текстів і побудови графів знань у сфері кібербезпеки. Проте проведений авторами аналіз здебільшого обмежується рівнем сутностей і відношень, тоді як питання формалізації зв'язків між тематичними кластерами, як окремого графового рівня, залишається недостатньо дослідженим.

Метод асоціативних правил традиційно використовують для виявлення статистично значущих залежностей у транзакційних даних. У дослідженні [5] автори показали, що асоціативні правила можуть застосовуватися для підтримки якісного аналізу та формування концептуальних моделей у складних інформаційних системах. Проте у роботі не розглядають їх інтеграції із графовими структурами знань.

У роботі [16] автори запропонували підхід, що розширює класичні асоціативні правила на графові структури (англ. *Graph Pattern Association Rules*). Автори довели можливість виявлення регулярностей безпосередньо у графа знаньх шляхом пошуку повторюваних підграфів.

Результатом роботи [11] є модель AR-KGAT, у якій асоціативні правила використовуються для покращення навчання подань у графа знаньх знань. Автори продемонстрували, що поєднання правил і графових механізмів уваги дає змогу враховувати високорівневі залежності.

Отже, у науковій літературі існують підходи до застосування асоціативних правил у графових структурах, однак вони або спрямовані на покращення подань, або на пошук підграфових закономірностей. Питання побудови окремого графа тематичних зв'язків на основі кластерів і правил та його подальшого структурного аналізу із використанням мережевих характеристик у контексті багатомовних кібербезпекових корпусів залишається відкритим.

Проведений аналіз останніх досліджень та публікацій показує, що сучасні дослідження забезпечують:

- автоматизоване вилучення сутностей і відношень із кібербезпекових текстів;
- побудову графів знань на рівні сутностей;
- окремі підходи до застосування асоціативних правил у графових структурах.

Водночас, у наукових працях відсутня формалізована модель, що поєднує двомовний граф знань, тематичну кластеризацію, асоціативні правила та мережеві метрики в межах окремого графа тематичних зв'язків із кількісним оцінюванням структурної важливості тематичних кластерів.

Матеріали та методи дослідження. У дослідженні використано корпус із 200 україномовних та 200 англомовних документів у сфері кібербезпеки. У дослідженні використано двомовний граф знань кібербезпеки, побудований у роботі [13], на основі якого здійснено семантичну кластеризацію сутностей та сформовано асоціативні правила за заданими пороговими значеннями показників *support*, *confidence* і *lift*. Отримані правила відображено на рівень тематичних кластерів для побудови орієнтованого зваженого графа тематичних зв'язків. Для аналізу структурної ролі кластерів застосовано мережеві метрики центральності та інтегрований показник, що поєднує топологічні й статистичні характеристики.

Результати дослідження та їх обговорення / Research results and their discussion

Формування двомовного графа знань, включно з процедурами вилучення сутностей, міжмовного узгодження та тематичної кластеризації, описано в попередньому дослідженні авторів [13]. У цій роботі запропоновано підхід до побудови двомовної графової моделі знань у сфері кібербезпеки на основі великих мовних моделей і багатомовних семантичних подань.

У межах цього дослідження отриманий граф знань використано як вихідну структуру для подальшої формалізації графа тематичних зв'язків.

Нехай

$$G = (V, E) \quad (1)$$

– орієнтований граф знань, де: V – множина сутностей, E – множина семантичних відношень між ними. Для ілюстрації структури такого подання графа наведемо його спрощений приклад. Нехай множина вершин містить сутності: v_1 – шкідливе ПЗ; v_2 – CVE-2026-1234; v_3 – Microsoft Office.

Тоді орієнтовані ребра можуть описувати семантичні відношення типу:

- $(v_1 \rightarrow v_2)$ – шкідливе ПЗ експлуатує вразливість;
- $(v_1 \rightarrow v_3)$ – шкідливе ПЗ націлене на програмний продукт.

У такому поданні вершини відповідають сутностям предметної області, а напрям ребра відображає семантику відношення між ними.

Семантична кластеризація вузлів графа знань дала можливість сформувати множину тематичних кластерів, а саме:

$$C = \{c_1, c_2, \ldots, c_k\}. \quad (2)$$

Асоціативні правила сформовано у попередньому дослідженні [13] на основі транзакційної моделі, в якій кожен документ корпусу розглядався як множина виявлених у ньому сутностей. У межах цієї роботи результати такого аналізу інтерпретовано на рівні тематичних кластерів шляхом відображення сутностей у відповідні кластери. Внаслідок цього було отримано множину асоціативних правил між тематичними кластерами, а саме:

$$AR = \{c_i \Rightarrow c_j \mid i, j = 1, \ldots, k\}, \quad (3)$$

де ліва частина правила c_i є антецедентом (англ. *Antecedent*), а права частина c_j є консеквентом (англ. *Consequent*). Кожне правило характеризується показниками *support*, *confidence* та *lift*. До множини AR внесено тільки ті правила, що задовольняють задані порогові значення цих показників. Множину зв'язків між кластерами визначено як

$$R = \{(c_i, c_j) \mid (c_i \Rightarrow c_j) \in AR\}. \quad (4)$$

Оскільки правила є спрямованими, то граф тематичних зв'язків також є орієнтованим. Вагомість кожного ребра визначають як

$$w(c_i, c_j) = Lift_{ij}(c_i \Rightarrow c_j). \quad (5)$$

У разі наявності правил у двох напрямах між однією парою кластерів формуються два орієнтованих ребра. Граф тематичних зв'язків визначають як

$$G_T = (C, R, w), \quad (6)$$

де: C – множина тематичних кластерів; R – множина орієнтованих зв'язків між кластерами; w – вагова функція. Для кращого розуміння виразу (6) наведемо спрощений приклад графа тематичних зв'язків. Нехай множина тематичних кластерів має вигляд $C = \{c_1, c_2, c_3\}$, де:

- c_1 – кластер "Шкідливе ПЗ та експлойти";
- c_2 – кластер "Вразливості маршрутизаторів";
- c_3 – кластер "Програмне забезпечення Oracle".

Припустимо, що за результатами асоціативного аналізу отримано правила $c_1 \rightarrow c_2$ зі значенням *lift* 2.10 та $c_2 \rightarrow c_3$ зі значенням *lift* 1.85. Тоді граф тематичних зв'язків G_T міститиме орієнтовані ребра $R = \{(c_1, c_2), (c_2, c_3)\}$, а вагова функція визначатиметься як:

- $w(c_1, c_2) = 2.10$;
- $w(c_2, c_3) = 1.85$.

Такий приклад показує, що вершини графа відповідають тематичним кластерам, орієнтовані ребра відображають напрям асоціативної залежності між темами, а їх вагомості характеризують силу міжтематичного зв'язку.

Для забезпечення коректного застосування мережевих метрик у роботі розглянуто два аспекти: врахування напряму зв'язків і врахування ваг ребер.

Оскільки асоціативні правила можуть існувати у двох напрямах для однієї пари кластерів, між вершинами c_i та c_j допускається наявність двох орієнтованих ребер (c_i, c_j) та (c_j, c_i) з різними вагами. Подальші розрахунки мережевих характеристик виконуються на орієнтованій структурі графа знань з урахуванням ваг ребер.

Для оцінювання ролі тематичних кластерів у структурі G_T використано дві мережеві характеристики. Центральність за посередництвом $CB(v)$ відображає, наскільки часто вершина v виступає "містком" на найкоротших шляхах між іншими вершинами, тобто характеризує її структурну роль у зв'язуванні різних тематичних областей. Власна векторна центральність $CE(v)$ відображає впливовість вершини v з урахуванням впливовості її сусідів; високі значення відповідають кластерам, які пов'язані з іншими "сильними" кластерами. Зазначені метрики розраховуються для всіх вершин $v \in C$ та використовуються як складові інтегрованого показника важливості. Обчислення зазначених метрик виконано для орієнтованого зваженого графа знань G_T з урахуванням ваг ребер, визначених за формулою (5).

Для узагальнення структурних і статистичних характеристик тематичних кластерів запропоновано інтегрований показник важливості ITS (англ. *Integrated Thematic Significance*):

$$ITS(v) = \alpha \cdot \bar{C}_B(v) + \beta \cdot \bar{C}_E(v) + \gamma \cdot \overline{MaxLift}(v), \quad (5)$$

де: $\overline{C_B}(v)$ – нормалізоване значення центральності за посередництвом; $\overline{C_E}(v)$ – нормалізоване значення центральності власного вектора; $\overline{MaxLift}(v)$ – нормалізоване максимальне значення *lift* для кластера *v*. Коефіцієнти α, β, γ мають значення більше або рівне 0, а їхня сума становить 1.

Кожна складова в інтегрованому показнику нормалізована за методом min-max нормалізації [7]:

$$\overline{x}(v) = \frac{x(v) - x_{\min}}{x_{\max} - x_{\min}}, \quad (6)$$

де: $\overline{x}(v)$ – нормоване значення ознаки, $x(v)$ – вхідне значення ознаки, x_{min} та x_{max} – відповідно мінімальне та максимальне значення даної ознаки у всьому наборі даних. Це забезпечує приведення значень до інтервалу [0, 1] та дає змогу поєднувати різного роду показники в межах однієї формули.

Табл. 1. Використані конфігурації параметрів інтегрованого показника важливості ITS / Parameter configurations used for the integrated importance measure ITS

α	β	γ	Опис показника	Суть показника
1	0	0	*ITS* визначають тільки структурним значенням кластера.	Оцінюється тільки "містковa" функція теми у графі – наскільки вона поєднує різні тематичні області.
0	1	0	*ITS* визначають тільки впливовістю кластера.	Визначає кластери, що пов'язані з іншими впливовими кластерами – тобто "ядро" тематичної структури.
0	0	1	*ITS* визначають тільки силою асоціативних правил.	Виділяються теми з найсильнішими та найбільш несподіваними статистичними зв'язками.
0,33	0,33	0,33	Рівноважна інтеграція структурного значення, впливовості та сили асоціативних зв'язків.	Базова нейтральна модель без пріоритету жодного з компонентів.
0,5	0,5	0	Перевага структурних характеристик графа знань без урахування *lift*-ефекту.	Модель орієнтована тільки на топологію. мережі, ігноруючи статистичну "несподіваність" зв'язків.
0,2	0,2	0,6	Домінування статистичної сили асоціативних правил за часткового врахування структури.	Модель акцентує на ризикових або нетривіальних зв'язках між темами.

Показник *MaxLift*(*v*) визначають як максимальне значення *lift* серед усіх асоціативних правил, у яких кластер *v* виступає як антецедент або консеквент.

У дослідженні використано корпус, що містить 200 англомовних і 200 україномовних документів у сфері кібербезпеки. Після побудови двомовного графа знань та застосування порогових значень для показників *support* $\geq$ 0,05, *confidence* $\geq$ 0,55 та *lift* $\geq$ 1,1 сформовано граф тематичних зв'язків, який містить k = 5 вузлів і m = 4 орієнтованих ребра.

У поточному експерименті використано різні конфігурації вагових коефіцієнтів *α, β, γ,* що визначає внесок структурних і статистичних характеристик у формування інтегрованого показника важливості (табл. 1). Ці конфігурації сформовано задля аналізу чутливості інтегрованого показника важливості *ITS* до різних інтерпретацій тематичної важливості:

- структурної з урахуванням мережевої топології;
- впливової з урахуванням зв'язків із сильними вузлами;
- статистичної з урахуванням інтенсивності асоціативних залежностей.

Такий підхід дає змогу оцінити стабільність ранжування кластерів за різних концептуальних підходів до визначення ризику.

Отримані результати можуть залежати від таких чинників:

- обмежений обсяг корпусу, всьо 400 документів, що впливає на стабільність виявлених асоціативних залежностей;
- вибір порогових значень *support*, *confidence* та *lift*, які визначають щільність графа тематичних зв'язків;
- використання min-max нормалізації, чутливої до крайніх значень показників;
- невелика кількість вузлів у графі, що може впливати на поведінку спектральних характеристик, зокрема центральності власного вектора.

Урахування зазначених факторів дає змогу інтерпретувати результати як експериментальні для поточного корпусу та передбачає можливість зміни структури графа знань й ранжування кластерів під час масштабування даних.

Аналіз результатів дослідження. На першому етапі побудовано граф тематичних зв'язків на основі асоціативних правил, отриманих для україномовного корпусу. Після застосування порогових значень *support*, *confidence* та *lift* сформовано орієнтований тематичний граф (рис. 1). Розмір вузла пропорційний значенню інтегрованого показника важливості *ITS*(*v*), що поєднує структурні характеристики графа знань і статистичну силу асоціативних зв'язків. Отже, більший вузол відповідає кластеру з вищою інтегральною оцінкою тематичної значущості у структурі графа знань.

Числа на ребрах відображають значення вагової функції $w(c_i, c_j)$, тобто значення показника *lift* відповідного асоціативного правила між кластерами. Більше значення *lift* означає сильнішу та менш випадкову статистичну залежність між тематичними областями. Товщина ребра пропорційна величині *lift* і візуально підсилює сприйняття сили зв'язку.

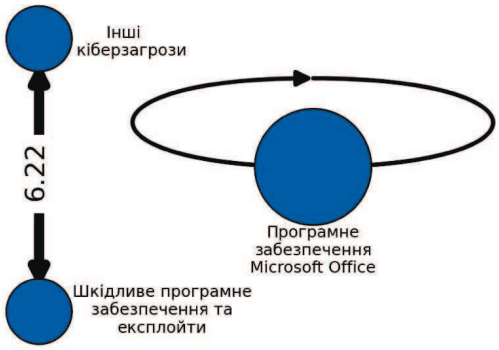

Рис. 1. Тематичний граф, сформований на основі україномовного корпусу / Thematic graph formed based on Ukrainian-language corpus

Отримана структура є відносно розрідженою та містить обмежену кількість міжкластерних зв'язків. Зокрема, спостерігається виражений орієнтований зв'язок між кластерами "Інші кіберзагрози" та "Шкідливе програмне забезпечення та експлойти". Водночас, кластер "Програмне забезпечення Microsoft Office" не формує значущих вихідних або вхідних зв'язків і поданий

ізольованим вузлом. Інші кластери мають слабку або відсутню взаємодію, тому вони не зображені на рис. 1.

На другому етапі побудовано граф тематичних зв'язків на основі двомовного графа знань, який об'єднує українські та англомовні дані. Формування тематичних кластерів здійснювалось у межах узгодженої багатомовної семантичної моделі. Після застосування тих самих порогових параметрів сформовано граф, що містить $k = 5$ вузлів і $m = 4$ орієнтованих ребра (рис. 2).

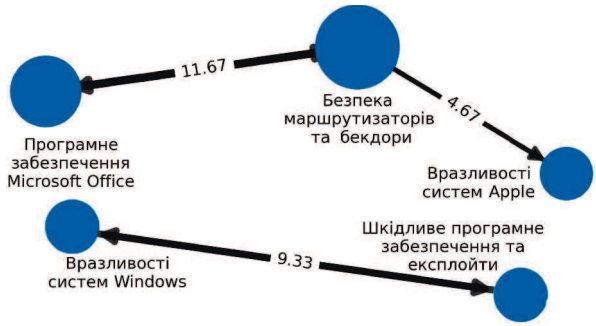

Рис. 2. Тематичний граф, сформований на основі двомовного корпусу / Thematic graph formed based on bilingual corpus

Структура двомовного графа знань є пов'язанішою та виражено ієрархічною. Кластер "Router Security and Backdoors" займає центральне положення у структурі графа знань. Разом з тим між кластерами "Microsoft Office Software" та "Router Security and Backdoors" зафіксовано найвищий рівень статистичної залежності. Кластер "Windows System Vulnerabilities" формує орієнтований зв'язок із кластером, пов'язаним зі шкідливим програмним забезпеченням та експлойтами. Водночас зв'язок між "Router Security and Backdoors" та "Apple System Vulnerabilities" є менш інтенсивним.

Візуалізація демонструє, що один із кластерів набуває ролі структурного центру, тоді як інші кластери формують послідовну лінійну конфігурацію залежностей.

Порівняння рис. 1 і 2 дає змогу виявити кілька принципових відмінностей.

По-перше, україномовний граф є розрідженішим. Кількість статистично значущих міжкластерних зв'язків є обмеженою, що зумовлено як розміром вибірки, так і специфікою тематичної структури корпусу. У межах україномовного аналізу сформовано тільки окремі спрямовані залежності, тоді як частина кластерів залишається ізольованою.

По-друге, у двомовному графі спостерігається зміна ролей окремих тематичних кластерів. Зокрема, кластер "Інші кіберзагрози", який формував виражений зв'язок в україномовному графі, у двомовній структурі не поданий як окремий активний вузол. Це можна пояснити тим, що після інтеграції англомовних даних частина тематичних залежностей перерозподіляється між специфічними кластерами, що отримують вищі значення *lift* і, відповідно, витісняють загальніші категорії з числа статистично значущих.

По-третє, хоча для англомовного корпусу окремо не було виявлено достатньо сильних асоціативних правил, його інтеграція у спільний граф знань істотно розширює тематичний контекст. Англомовні кластери не формують самостійної щільної структури, однак вони підсилюють вже наявні тематичні залежності через міжмовні узгодження та спільні семантичні області. Внаслідок цього формується пов'язана і структурно виражена конфігурація.

Отже, інтеграція різномовних сегментів у двомовний граф:

- змінює вагові характеристики зв'язків;
- впливає на ранжування тематичних кластерів за показником *ITS*;
- дає змогу виявити структурний центр, який в одномовному аналізі не був настільки очевидним.

Також встановлено, що двомовному графу тематичних зв'язків характерна вираженіша централізація структури, у межах якої один із кластерів набуває ролі вузла-концентратора, тоді як інші теми формують спрямований ланцюг залежностей. Це свідчить про те, що розширення корпусу за рахунок іншої мови підвищує стабільність тематичної структури та дає змогу уточнити характер міжкластерних взаємодій.

Табл. 2. Ранжування тематичних кластерів за різних конфігурацій параметрів ITS / Ranking of thematic clusters under different ITS parameter configurations

(α, β, γ)	1	2	3	4	5
(1, 0, 0)	en_1	en_0	uk_2	en_2	uk_4
(0, 1, 0)	en_1	uk_2	en_0	en_2	uk_4
(0, 0, 1)	en_1	uk_2	en_0	en_2	uk_4
(0.33, 0.33, 0.33)	en_1	uk_2	en_0	en_2	uk_4
(0.5, 0.5, 0)	en_1	uk_2	en_0	en_2	uk_4
(0.2, 0.2, 0.6)	en_1	uk_2	en_0	en_2	uk_4

Відповідність позначень кластерів:

- en_1 – Router Security and Backdoors;
- uk_2 – шкідливе програмне забезпечення та експлойти;
- en_0 – Microsoft Office Software;
- uk_4 – інші кіберзагрози;
- en_2 – програмне забезпечення Microsoft Office.

Після побудови графа тематичних зв'язків для двомовного корпусу проведено оцінювання інтегрованого показника важливості *ITS* за шістьма конфігураціями вагових коефіцієнтів α, β, γ, наведеними в табл. 1. Для кожної конфігурації сформовано ранжування кластерів за спаданням *ITS*. Відповідні результати наведено в табл. 2.

Аналіз отриманих результатів показує, що незалежно від конфігурації параметрів α, β, γ склад топ-5 кластерів залишається незмінним. Зміни вагових коефіцієнтів практично не впливають на порядок ранжування, за винятком незначної перестановки другого та третього місця у конфігурації (1,0,0), де враховується винятково структурна центральність за посередництвом.

Такий результат має декілька важливих інтерпретацій. По-перше, домінантні кластери демонструють узгоджено високі значення одразу за трьома компонентами інтегрованого показника важливості – структурним значенням, впливовістю у мережі та максимальною статистичною силою асоціативних залежностей. Це означає, що вони є водночас структурними "містками" між тематичними областями, пов'язаними з іншими впливовими вузлами, а також учасниками найсильніших статистичних залежностей.

По-друге, стабільність ранжування свідчить про внутрішню узгодженість структури тематичного графа знань. Структурна топологія та статистична сила зв'язків не конфліктують між собою, а взаємно підсилюють виділення одних і тих самих тематичних центрів.

По-третє, нечутливість *ITS* до варіації ваг може бути пов'язана з обмеженим обсягом корпусу. За відносно невеликої кількості тематичних кластерів і зв'язків домінантні теми природно отримують високі значення за всіма компонентами показника одночасно. У масштаб-

нішому корпусі можна очікувати значно більшої диференціації ролей та, відповідно, більшої чутливості до параметрів α, β, γ.

Отже, проведений аналіз результатів дослідження підтверджує коректність інтегрованого підходу до оцінювання тематичної важливості кластерів у графі тематичних зв'язків та демонструє, що ключові кластери зберігають провідні позиції незалежно від концептуальної інтерпретації ризику – структурної, впливової чи статистичної.

Обговорення результатів дослідження. У роботі [12] здійснено систематизацію сучасних підходів до застосування графів знань у розвідці кіберзагроз. Автори показують, що переважна більшість досліджень зосереджена на автоматизованому наповненні графів знань, інтеграції різнорідних джерел та прикладних сценаріях аналітики, зокрема підтримці виявлення інцидентів. Водночас міжтематичний рівень залежностей у таких роботах, зазвичай, не формалізується як окремий аналітичний шар. Це підтверджує актуальність підходу, запропонованого у цьому дослідженні, де акцент зроблено саме на побудові орієнтованого зваженого графа тематичних зв'язків.

У роботі [14] розглянуто задачу видобування асоціативних правил шляху у великих графах властивостей. Автори аналізують закономірності, що виникають у структурі вже сформованого графа, і досліджують правила, які описують повторювані патерни на його шляхах. На відміну від цього підходу, у даному дослідженні асоціативні правила виводяться безпосередньо з текстового корпусу, після чого агрегуються на міжкластерному рівні та використовуються для формування окремого тематичного графа знань. Отже, правила виступають не лише інтерпретованими патернами, а структуроутворювальним механізмом для побудови нового рівня узагальнення.

У роботі [15] запропоновано інтеграцію характеристик центральності у процес формування ознак для завдань виявлення вторгнень у мережах IoT. Центральності розглядаються як індикатори важливості вузлів у мережевій структурі. Хоча прикладний фокус цієї роботи пов'язаний із завдання кібербезпеки на рівні мережевого трафіку, сама ідея використання топологічних характеристик як показників впливовості узгоджується з логікою цього дослідження. Водночас у цій роботі центральності обчислюються на міжтематичному графі, побудованому з асоціативних правил, а інтегрований показник поєднує структурні характеристики з максимальним значенням *lift*, що дає змогу враховувати як топологічну роль кластера, так і статистичну силу його зв'язків.

У роботі [2] запропоновано методологію, що поєднує графи знань із великими мовними моделями для підвищення пояснюваності та якості аналітики. Автори інтегрують графове представлення знань із компонентами інтерпретованих штучних інтелектуальних систем та LLM-генеруванням звітів для пояснюваного виявлення загроз. На відміну від такого прикладного фокусу, це дослідження не оптимізує детекцію інцидентів і не спрямоване на автоматичну генерацію аналітичних звітів, а формалізує міжтематичні залежності як окремий рівень структурування знань і забезпечує кількісне ранжування кластерів за їх структурною роллю.

У роботі [17] узагальнено сучасні підходи до міжмовного зіставлення сутностей на основі представлень та виділено ключові модулі, що впливають на якість узгодження, зокрема механізми агрегації структурної та семантичної інформації. Це дає змогу розглядати інтеграцію україномовного та англомовного корпусів у межах цього дослідження як закономірний етап підсилення семантичних зв'язків. Отримані результати свідчать, що об'єднання двох мовних підмножин сприяє виявленню стабільніших міжкластерних залежностей, що не завжди проявляються в окремому монолінгвальному аналізі.

У роботі [1] описано автономне формування кібербезпекових графів знань шляхом інтеграції гетерогенних джерел та подальшого застосування правил і графових алгоритмів для посилення аналітичних можливостей. Проте навіть у таких підходах основна увага приділяється побудові самого графа знань і його прикладному використанню. У цьому дослідженні акцент зміщено на наступний рівень узагальнення – формування окремого графа тематичних зв'язків та введення кількісного механізму ранжування, який дає змогу оцінити структурну важливість тем незалежно від конкретного сценарію застосування.

Отже, проведене обговорення результатів дослідження показало, що хоча сучасні роботи активно розвивають напрями побудови графів знань, автоматизованого вилучення знань та інтеграції LLM у кібербезпекову аналітику, міжтематичний рівень залежностей як окремий об'єкт структурного аналізу представлений обмежено. Це підтверджує доцільність запропонованого підходу, у межах якого асоціативні правила, семантична кластеризація та мережеві метрики інтегруються для формування показника тематичної важливості. Отримані результати засвідчують стабільність провідних кластерів за різних конфігурацій оцінювання, що обґрунтовує новизну та практичну значущість проведеного дослідження.

Отже, внаслідок виконаної роботи можна сформулювати такі наукову новизну та практичну значущість результатів дослідження.

Наукова новизна отриманих результатів дослідження – розроблено метод формалізації графа тематичних зв'язків як окремого рівня подання знань над двомовним графом знань у сфері кібербезпеки, який забезпечує перехід від аналізу локальних асоціативних залежностей між сутностями до кількісного оцінювання структурної ролі тематичних кластерів на основі інтеграції мережевих метрик та статистичних характеристик асоціативних правил у межах багатомовної моделі.

Практична значущість результатів дослідження – можна застосувати запропонований метод формалізації графа тематичних зв'язків для структурного аналізу тематичних напрямів у сфері кібербезпеки, виявлення міжтематичних залежностей у багатомовному інформаційному середовищі та під час пріоритизації напрямів моніторингу потенційно критичних загроз.

Висновки / Conclusions

Розроблено підхід до формалізації графа тематичних зв'язків як окремого рівня подання знань над двомовним графом знань у сфері кібербезпеки, який дав можливість перейти від аналізу локальних асоціативних залежностей між сутностями до структурного оцінювання ролі тематичних кластерів у межах інтегрованої

багатомовної моделі. За результатами проведеного дослідження можна зробити такі основні висновки.

1. Формалізовано граф тематичних зв'язків як орієнтовану зважену структуру, побудовану на основі асоціативних правил між тематичними кластерами що отримані внаслідок аналізу двомовного текстового корпусу із застосуванням показників *support*, *confidence* та *lift*, що дало змогу перейти від аналізу локальних зв'язків між окремими сутностями до структурного подання міжтематичних залежностей та дослідити їх як самостійний об'єкт графового аналізу.
2. Запропоновано інтегрований показник важливості *ITS*, який поєднує мережеві характеристики та статистичну силу асоціативних залежностей, що забезпечує кількісне оцінювання структурної важливості тематичних кластерів і дає змогу ранжувати їх за ступенем впливовості у межах тематичного графа.
3. Встановлено стабільність ранжування кластерів за різних конфігурацій вагових коефіцієнтів α, β, γ шляхом варіювання параметрів інтегрованого показника та повторного обчислення позицій кластерів, що свідчить про узгодженість структурних і статистичних характеристик домінантних тематичних напрямів у сфері кібербезпеки та підтверджує надійність запропонованого критерію оцінювання.
4. Показано вплив міжмовної інтеграції шляхом порівняльного аналізу україномовної та двомовної моделей графа тематичних зв'язків, що забезпечує формування більш пов'язаної структури залежностей та посилення міжкластерних зв'язків і демонструє підвищення структурної цілісності моделі порівняно з одномовним представленням.
5. Сформовано напрями подальших етапів дослідження, що передбачають розширення обсягу корпусу даних, аналіз чутливості інтегрованого показника важливості на більших тематичних структурах, а також дослідження динаміки тематичних залежностей у часі та їх використання для прогнозування еволюції кіберзагроз.

References

1. Alharbi, H., Hur, A., Alkahtani, H., & Ahmad, H. F. (2025). Enhancing cybersecurity through autonomous knowledge graph construction by integrating heterogeneous data sources. *PeerJ Computer Science*, 11, article ID e2768. https://doi.org/10.7717/peerj-cs.2768
2. Belcastro, L., Carlucci, C., Cosentino, C., Liò, P., & Marozzo, F. (2026). Enhancing network security using knowledge graphs and large language models for explainable threat detection. *Future Generation Computer Systems*, 176, article ID 108160. https://doi.org/10.1016/j.future.2025.108160
3. Bratsas, C., Anastasiadis, E. K., Angelidis, A. K., Ioannidis, L., Kotsakis, R., & Ougiaroglou, S. (2024). Knowledge Graphs and Semantic Web Tools in Cyber Threat Intelligence: A Systematic Literature Review. *Journal of Cybersecurity and Privacy*, *4*(3), 518–545. https://doi.org/10.3390/jcp4030025
4. Cheng, Y., Bajaber, O., Tsegai, S. A., Song, D., & Gao, P. (2025). CTINexus: Automatic Cyber Threat Intelligence Knowledge Graph Construction Using Large Language Models. arXiv preprint arXiv:2410.21060. https://doi.org/10.48550/arXiv.2410.21060
5. Fan, W., Wang, X., Wu, Y., & Xu, J. (2015). Association rules with graph patterns. *Proceedings of the VLDB Endowment*, *8*(12), 1502–1513. https://doi.org/10.14778/2824032.2824048
6. Fieblinger, R., Alam, M. T., & Rastogi, N. (2024). Actionable Cyber Threat Intelligence using Knowledge Graphs and Large Language Models. arXiv preprint arXiv:2407.02528. https://doi.org/10.48550/arXiv.2407.02528
7. Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Burlington: Morgan Kaufmann. URL: https://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf
8. Hu, Y., Zou, F., Han, J., Sun, X., & Wang, Y. (2024). LLM-TIKG: Threat intelligence knowledge graph construction utilizing large language model. *Computers & Security*, 145, article ID 103999. https://doi.org/10.1016/j.cose.2024.103999
9. Huang, L., & Xiao, X. (2024). CTIKG: LLM-Powered Knowledge Graph Construction from Cyber Threat Intelligence. *First Conference on Language Modeling*. URL: https://openreview.net/forum?id=DOMP5AqwQz
10. Li, Y., Thomas, M., & Osei-Bryson, K. M. (2014). Using Association Rules Mining to Facilitate Qualitative Data Analysis in Theory Building. In *Lecture Notes in Information Systems and Organisation, 34*, 79–91. https://doi.org/10.1007/978-1-4614-9463-8_7
11. Mouiche, I., & Saad, S. (2025). Entity and relation extractions for threat intelligence knowledge graphs. *Computers & Security*, 148, article ID 104120. https://doi.org/10.1016/j.cose.2024.104120
12. Papoutsoglou, M., Meditskos, G., Bassiliades, N., Kontopoulos, E., & Vrochidis, S. (2024). Mapping the Current Status of CTI Knowledge Graphs through a Bibliometric Analysis. In *SETN 2024*. ACM. https://doi.org/10.1145/3688671.3688738
13. Pavlyshenko, B., Stasiuk, M. (2025). Multilingual Knowledge Graph for Thematic Text Analysis. In *International Scientific and Technical Conference Security of Modern Information and Communication Systems* (SMICS-2025), pp. 380–384. URL:. https://smics.lnu.edu.ua/uk/zbirnyk/
14. Sasaki, Y., & Karras, P. (2024). Mining Path Association Rules in Large Property Graphs. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (pp. 1994–2003). https://doi.org/10.1145/3627673.3679525
15. Termos, M., Ghalmane, Z., Brahmia, M. E. A., Fadlallah, A., Jaber, A., & Zghal, M. (2024). GDLC: A new Graph Deep Learning framework based on centrality measures for intrusion detection in IoT networks. *Internet of Things*, 26, article ID 101214. https://doi.org/10.1016/j.iot.2024.101214
16. Zhang, Z., Huang, J., & Tan, Q. (2020). Association Rules Enhanced Knowledge Graph Attention Network. arXiv preprint arXiv:2011.08431. https://doi.org/10.48550/arXiv.2011.08431
17. Zhu, B., Wang, R., Wang, J., Shao, F., & Wang, K. (2024). A survey: knowledge graph entity alignment research based on graph embedding. *Artificial Intelligence Review*, *57*(9), article ID 229. https://doi.org/10.1007/s10462-024-10866-4

B. M. Pavlyshenko, M. I. Stasiuk

Ivan Franko National University of Lviv, Lviv, Ukraine

ANALYSIS OF THEMATIC RELATIONSHIPS IN A MULTILINGUAL CYBERSECURITY KNOWLEDGE GRAPH

This study extends methodology for analyzing cybersecurity text corpora by formalizing inter-theme dependencies as a distinct structural layer above a bilingual knowledge graph, thereby enabling a transition from the analysis of individual entities to a network-based interpretation of thematic interactions. A bilingual knowledge graph was constructed based on 200 Ukrainian and 200 English documents; semantic clustering of entities was performed, and association rules were generated using predefined thresholds for *support*, *confidence*, and *lift*. The resulting rules were projected onto thematic clusters to construct a directed weighted thematic relationship graph in which *lift* serves as the edge weight. It is demonstrated that preserving directionality is essential, since ignoring edge

orientation leads to the loss of information about the configuration of thematic flows. As a result, the directed weighted thematic relationship graph enables the representation of cross-theme influence patterns. The Integrated Thematic Significance (*ITS*) index is introduced, combining betweenness centrality, eigenvector centrality, and the maximum *lift* associated with each cluster, and its sensitivity was analyzed across different weighting coefficient configurations. The ranking of dominant clusters remains stable under all tested parameter settings, indicating consistency between structural topology and statistical association strength. In particular, clusters related to router security, malware and exploits, and office software simultaneously function as structural bridges and participate in the strongest associative dependencies. Moreover, the use of two languages resulted in the formation of a denser and more hierarchically organized thematic relationship graph structure compared to the Ukrainian-only model, even though the English subset alone does not generate sufficiently strong rules. Consequently, the integration of Ukrainian and English corpora in the graph construction process reinforced latent inter-cluster dependencies that did not reach threshold values in the monolingual corpus and made it possible to identify a structural hub within the graph. The advantages of the research results consist in the formalization of a cluster-level dependency model and the development of a quantitative ranking mechanism that can be used for analytical interpretation and identification of structurally dominant thematic domains in cybersecurity. Future research will focus on corpus scaling, temporal analysis of thematic flows, and robustness evaluation of the integrated index in large-scale networks.

Keywords: association rules; semantic clustering; cross-lingual alignment; network metrics; information structuring.