КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 3 (27), 2025
ISSN 2663 - 4023
DOI 10.28925/2663-4023.2025.27.712
УДК 004.912:004.942

Пучков Олександр Олександрович
кандидат філософських наук, професор
начальник Інституту спеціального зв'язку та захисту інформації
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна
ORCID ID: 0000-0002-8585-1044
iszzi@iszzi.kpi.ua

Ланде Дмитро Володимирович
доктор технічних наук, професор, завідувач кафедри
Навчальний фізико-технічний інститут
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна
ORCID ID: 0000-0003-3945-1178
dwlande@gmail.com

Субач Ігор Юрійович
доктор технічних наук, професор, завідувач кафедри
Інститут спеціального зв'язку та захисту інформації
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна
ORCID ID: 0000-0002-9344-713X
igor_subach@ukr.net

ЗАСТОСУВАННЯ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ ДЛЯ ПОБУДОВИ «ЛІСУ ІЄРАХІЙ ТЕРМІНІВ»

Анотація. Одним із способів упорядкування та систематизації знань є формування термінологічних онтологій, які дозволяють структурувати інформацію в конкретних предметних областях, таких як кібербезпека. У зв'язку з революційною появою великих мовних моделей (large language model, LLM) з'являються нові можливості для автоматизації процесу побудови «лісу ієрархій термінів» (ЛІТ). Побудова ЛІТ є необхідною для таких кількох ключових аспектів у сфері кібербезпеки та управління знаннями, як уніфікація термінології, покращення комунікації, оптимізація інформаційного пошуку, систематизація знань, адаптація до нових викликів, підтримка досліджень та інновацій. У статті розглядається роль LLM у побудові ЛІТ в контексті сучасних викликів інформаційного середовища. Завдяки революційним досягненням у сфері штучного інтелекту, LLM забезпечують автоматизацію та оптимізацію процесів обробки, аналізу та структурування великих обсягів текстових даних. Описано ключові етапи реалізації ЛІТ за допомогою LLM, зокрема обробка текстових даних, визначення дискримінантної сили термінів, встановлення зв'язків між ними та візуалізація результатів. Запропоновано методику визначення асоціативних зв'язків між заздалегідь визначеними термінами для побудови ЛІТ. Наведено приклади практичної реалізації запропонованої методики на основі застосування інформаційно-аналітичної системи «Кібер Агрегатор». Продемонстровано приклад формування промпту для побудови ЛІТ до системи генеративного штучного інтелекту DeepSeek.com. Запропоновано технологію візуалізації ЛІТ шляхом застосування програми для аналізу і візуалізації графів CSV2Graph. Використання запропонованих технологій дозволяє підвищити ефективність і точність побудови термінологічних онтологій, що є важливим для адаптації до швидко зростаючих інформаційних потоків у сучасному світі.

Ключові слова: великі мовні моделі; штучний інтелект; ліс ієрархій термінів; термінологічні онтології; візуалізація даних; кібербезпека.

ВСТУП

У сучасному світі інформаційні технології та кібербезпека набувають дедалі більшої значущості. Зі зростанням обсягу неструктурованих даних та динамічних інформаційних потоків виникає потреба в ефективних методах їх обробки, пошуку та навігації. Одним із способів упорядкування та систематизації знань є формування термінологічних онтологій, які дозволяють структурувати інформацію в конкретних предметних областях, таких як кібербезпека. У цьому контексті особливу увагу слід приділити розробці «лісу ієрархій термінів» (ЛІТ), який базується на концептно-об'єктних зв'язках.

Постановка проблеми. У зв'зку з революційною появою LLM з'являються нові можливості для автоматизації процесу побудови ЛІТ. Побудова ЛІТ є необхідною для таких кількох ключових аспектів у сфері кібербезпеки та управління знаннями, як уніфікація термінологій, покращення комунікації, оптимізація інформаційного пошуку, систематизація знань, адаптація до нових викликів, підтримка досліджень та інновацій.

У теперішній час у сфері кібербезпеки існує множина термінів та концептів, які можуть мати різні значення в залежності від контексту, проте ЛІТ дозволяє створити стандартизовану термінологію, що сприяє зменшенню непорозумінь між фахівцями, які працюють у цій галузі.

Зрозуміле спілкування є критично важливим для ефективного співробітництва між різними командами та організаціями. ЛІТ допомагає структурувати знання та ідеї, забезпечуючи чітке спілкування між фахівцями з різних областей, що, в свою чергу, сприяє більш швидкому вирішенню проблем. Крім того, ЛІТ дозволяє організувати знання в структурованій формі, що допомагає в управлінні великими обсягами інформації. Це забезпечує більш ефективне навчання та підготовку фахівців у сфері кібербезпеки.

Оскільки технології та загрози в сфері кібербезпеки постійно еволюціонують, ЛІТ забезпечує гнучкість в адаптації термінології до нових умов. Це дозволяє фахівцям швидше реагувати на нові загрози та розробляти адекватні стратегії їхнього подолання. Завдяки впровадженню ЛІТ, процес пошуку та доступу до інформації стає більш ефективним. Відповідні терміни та їх ієрархії можуть бути використані для покращення алгоритмів пошуку, що призводить до більш швидкого та точного знаходження необхідних даних. Створення ЛІТ слугує основою для проведення наукових досліджень у сфері кібербезпеки. Чітка термінологія сприяє кращому розумінню існуючих проблем та розробці нових рішень, що стимулює інновації у цій сфері.

Таким чином, побудова ЛІТ є важливим кроком для підвищення ефективності роботи у сфері кібербезпеки та сприяє створенню єдиного інформаційного простору, в якому терміни та концепти взаємодіють у систематизованій формі.

Аналіз останніх досліджень і публікацій. Розвиток інформаційних технологій та зростання загроз у сфері кібербезпеки вимагають ефективного управління термінами та онтологіями, що виникають у цій динамічній галузі. У зв'язку з цим, побудова ЛІТ стає важливим напрямом досліджень. Багато науковців досліджували формування термінологічних онтологій, що сприяють створенню чіткої та уніфікованої мови у сфері кібербезпеки.

Згідно з роботою [1], онтології є критично важливими для формалізації знань у будь-якій доменній області, включаючи кібербезпеку. Автори пропонують рішення у вигляді управління знаннями на основі онтології, яке забезпечує спільне розуміння, інтеграцію і обмін даними в середовищі Cyber Defense Exercises (CDX). Для цього, за

допомогою мов RDF (Resource Description Framework) і OWL (Web Ontology Language), була побудована онтологія CDX, що організує дані у форматі, зрозумілому для машин, яка може бути корисною для обчислювальної лінгвістики.

Подібні ідеї розвиваються у технічному звіті [2], де основна увага акцентується на інформаційному пошуку та онтологіях у сфері кібербезпеки, які створені для полегшення організації та управління знаннями в ній.

У статті [3] обговорюється специфіка термінології в кібербезпеці та пропонуються підходи до формалізації термінологій через онтології. Дана робота присвячена аналізу термінології, пов'язаної з кіберпростором, і зокрема, використанню терміну «cyber» у законодавстві Європейського Союзу та його перекладу на інші мови Це дослідження підкреслює необхідність стандартизації термінів для підвищення ефективності обміну інформацією та зменшення непорозумінь у сфері кібербезпеки.

У дослідженні [4] аналізуються виклики та можливості, що виникають при розробці онтологій у сфері кібербезпеки критичної інфраструктури і кіберфізичних систем. Автори пропонують розробити **інтегровану онтологію**, яка поєднує кібер- і фізичну безпеку та методи оцінки корисності створеної онтології для досягнення мети інтероперабельності.

Автори роботи [5] акцентують увагу на ролі технології Text Mining в екстракції інформації з кібербезпекових даних, що сприяє формуванню онтологій. Стаття стосується розробки системи **CASIE**, призначеної для автоматичного вилучення інформації про події, пов'язані з кібербезпекою, з текстів та її подальшої імплементації до семантичної моделі, яку можна інтегрувати у граф знань з кібербезпеки.

Дослідження [6] присвячене огляду графових моделей даних у сфері кібербезпеки, зокрема, використанню графів знань для обробки великих обсягів складних даних з різних джерел. У статті підкреслюється важливість графових моделей даних, як потужного інструменту для покращення аналізу та управління знаннями в області кібербезпеки, що може суттєво підвищити ефективність реагування на кіберзагрози.

У статтях [7], [8] пропонується методика виявлення і побудови мереж ієрархій термінів на основі аналізу текстових корпусів відповідної тематики. Методика базується на застосуванні методології компактифікованих графів горизонтальної видимості [9].

Наведений огляд літератури свідчить про те, що хоча вже зроблено багато для розробки термінологічних онтологій у сфері кібербезпеки, існує потреба у нових методологічних підходах, які б враховували специфіку даної галузі та інтегрували сучасні технології, такі як LLM, для створення більш адаптивних та ефективних систем управління термінами.

Мета дослідження. Метою даної статті є розробка методики формування ЛІТ, що дозволяє визначати змістовні зв'язки між термінами та ключовими концептами в області кібербезпеки.

Для досягнення мети вирішуються наступні взаємопов'язані завдання:

1. Розробка покрокової методики формування ЛІТ на основі текстових документів, що стосуються сфери кібербезпеки.
2. Визначення ролі LLM у процесі ідентифікації та оцінки термінів, їх зв'язків і значень.
3. Реалізація практичного застосування запропонованої методики на основі інформаційно-аналітичних систем, на прикладі системи «Кібер Агрегатор» [10].

ISSN 2663 - 4023

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Методику визначення асоціативних зв'язків між заздалегідь визначеними термінами (відношення «загальне-часткове») для побудови ЛІТ, можна розглядати як основу для формування деякої моделі предметної області. ЛІТ базується на сутностях із тексту, ключових словах і словосполученнях, методологію виявлення яких наведено у [5] – [8]. Використання ЛІТ дозволяє формувати пошукові образи, зокрема, при обробці інформації за тематикою кібербезпека.

При побудові ЛІТ передбачається відмова при її формуванні від спеціальних семантичних методик. Усі зв'язки в такій мережі визначаються природним застосуванням слів і словосполучень, що екстрагуються з репрезентативних текстових корпусів. ЛІТ, що має формуватися повністю автоматично, може розглядатися як основа для подальшого автоматизованого формування моделі предметної області.

Методика формування ЛІТ передбачає реалізацію послідовності кроків, що охоплюють попередню обробку вихідного текстового корпусу, вибір і сортування найбільш вагомих термінів, безпосереднє формування ЛІТ та його відображення.

Формальна постановка задачі виглядає наступним чином.

Дано: $D = \{d_1, d_2, \ldots, d_n\}$ — множина документів текстового корпусу з тематики кібербезпека;

$T = \{t_1, t_2, \ldots, t_m\}$ — множина термінів (слів), що зустрічаються у корпусі D;

$B = \{b_1, b_2, \ldots, b_k\}$ — множина біграм (словосполучень з двох слів), що зустрічаються в корпусі D;

$Tr = \{tr_1, tr_2, \ldots, tr_l\}$ — множина триграм (словосполучень з трьох слів), що зустрічаються в корпусі D;

$F(t_i)$ — частота терміну t_i у корпусі D;

$DF(t_i)$ — кількість документів, в яких зустрічається термін t_i;

$TF(t_i, d_j)$ — кількість появ терміну t_i у документі d_j;

$IDF(t_i) = \log \frac{N}{DF(t_i)}$ — обернена частота документів для терміну t_i;

$TF \cdot IDF(t_i, d_j) = TF(t_i, d_j) \cdot IDF(t_i)$ — значення *TF-IDF* для терміну t_i у документі D.

Необхідно: сформувати трирівневу мережу, в якій вузли представляють терміни, а зв'язки відповідають входженням одних термінів до інших.

Для вирішення сформульованої задачі необхідно виконати наступні кроки.

Крок 1. Попередня обробка текстових даних. В якості вхідних текстових корпусів з тематики кібербезпека обирається корпус повідомлень із соціальних медіа, отриманих, наприклад, за допомогою системи «Кібер Агрегатор» [10], шляхом опрацювання сформованого до нього запиту, наприклад, «Кібератака на Україну» *or* «Кібератака проти України» за першу половину 2022 року, обсягом понад 5000 знайдених документів із вебсайтів і соціальних мереж.

ISSN 2663 - 4023

Попередня обробка отриманих текстових корпусів передбачає вилучення флексій слів, що входять до цього корпусу, а також вилучення нетекстових символів тощо. Крім того, подальшому аналізу не підлягають терміни, що входять до, так званого, стоп-словника.

Крок 2. Оцінка дискримінантної сили для окремих слів. Кожному окремому слову з текстового масиву, що аналізується, ставиться у відповідність оцінка його «дискримінантної сили». Для цього для кожного терміну $t_i \in T$ обчислюється значення *TF-IDF*:

$$TF \cdot IDF\left(t_i\right) = \frac{1}{n}\sum_{j=1}^{n} TF\left(t_i, d_j\right) \cdot IDF\left(t_i, d_j\right), \quad (1)$$

де $TF\left(t_i, d_j\right)$ — частота появи біграми t_i у документі d_j.

Крок 3. Виконується те ж саме, що і на попередньому кроці, тільки для словосполучень із двох слів (біграм). Для кожної біграми $b_k \in B$ обчислюється значення *TF-IDF*:

$$TF \cdot IDF\left(b_k\right) = \frac{1}{n}\sum_{j=1}^{n} TF\left(b_k, d_j\right) \cdot IDF\left(b_k, d_j\right), \quad (2)$$

де $TF\left(b_k, d_j\right)$ — частота появи біграми b_k у документі d_j.

Крок 4. Виконуються ті ж самі дії, що і на попередньому кроці, тільки для словосполучень із трьох слів (триграм). Для кожної триграми $tr_l \in Tr$ обчислюється значення *TF-IDF* :

$$TF \cdot IDF\left(tr_l\right) = \frac{1}{n}\sum_{j=1}^{n} TF\left(tr_l, d_j\right) \cdot IDF\left(tr_l, d_j\right), \quad (3)$$

де $TF\left(tr_l, d_j\right)$ — частота появи триграми tr_l у документі d_j.

Зауважимо, що LLM можуть оцінювати значення термінів у контексті, використовуючи методи, такі як TF-IDF, що дозволяє виділити найзначніші слова та словосполучення для подальшої побудови ЛІТ.

Крок 5. Експертним методом визначається необхідний обсяг ЛІТ (число N), після чого обирається відповідна кількість окремих слів, біграм і триграм (всього $N \times N \times N$ елементів) з найбільшими ваговими значеннями. З відібраних на попередньому кроці елементів будується ЛІТ, в якому в якості вузлів розглядаються самі терміни, а зв’язки відповідають входженням одних термінів до інших.

У мережі, яка представляє ЛІТ, першому рівню відповідає вибрана множина одиничних слів, другому — множина біграм, а третьому — множина триграм. Якщо одиничне слово входить до біграми або триграми, або біграма входить до триграми, утворюється зв’язок, який позначається стрілкою. Множина вузлів, яким відповідають терміни і зв’язки утворює трирівневу мережу ЛІТ (див. рис. 1).

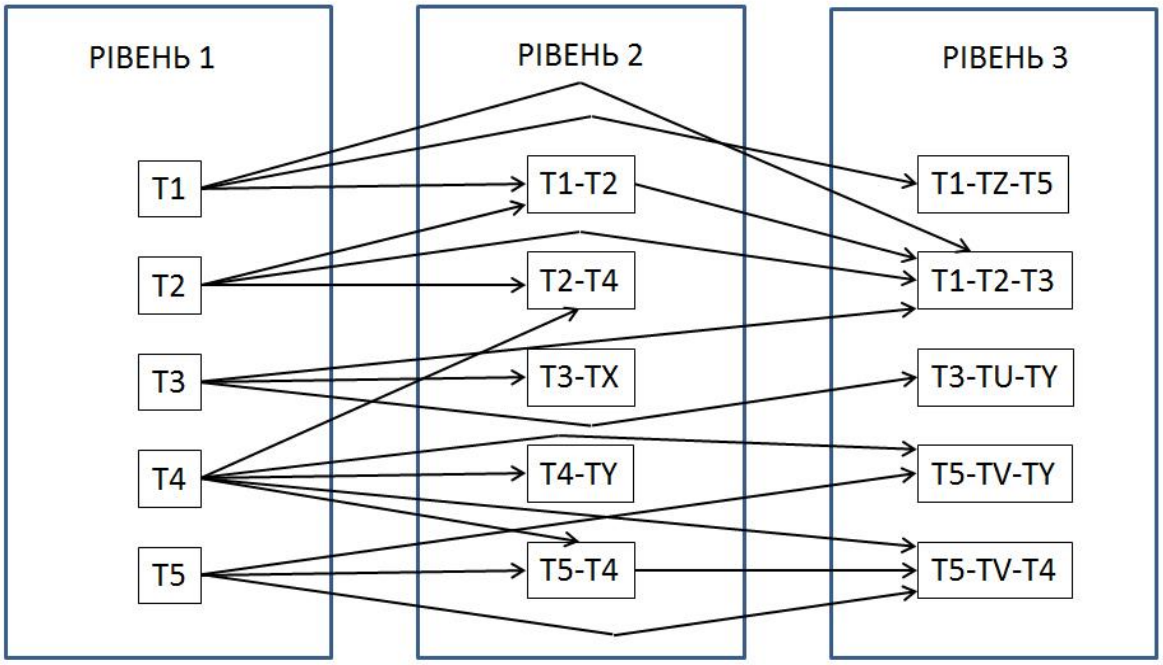

Рис. 1. Формування зв'язків у трирівневому ЛІТ

Крок 6. На останньому етапі формування ЛІТ здійснюється його відображення за допомогою програми аналізу і візуалізації графів, наприклад, CSV2Graph [11], в якій, для завантаження мереж природних ієрархій термінів до бази даних цієї системи достатньо привести множину термінів і визначених зв'язків до формату *csv*.

Сьогодні побудова ЛІТ стала можливішою завдяки революційній появі LLM та технологій штучного інтелекту. Ці інструменти забезпечують новий рівень обробки, аналізу та структурування великих обсягів інформації, що особливо важливо в умовах стрімкого зростання неструктурованих даних, зокрема у сфері кібербезпеки.

Завдяки LLM, процес формування ЛІТ може бути автоматизованим і оптимізованим. Інформаційні технології, що використовують дані моделі здатні здійснювати попередню обробку текстів, виокремлювати ключові терміни та словосполучення, а також встановлювати асоціативні зв'язки між ними. Використання таких технологій дозволяє значно підвищити ефективність та точність побудови термінологічних онтологій.

Основні етапи реалізації ЛІТ із застосуванням LLM включають **обробку текстових даних, визначення «ваги», інформативності термінів, встановлення зв'язків між термінами, візуалізацію результатів.**

ISSN 2663 - 4023

Виберіть із тексту найбільш інформативні стійки словосполучення-сутності, що відповідають поняттям з кібербезпеки довжиною точно у три слова (3-грами). Триграми не мають містити 4-х слів. Наприклад, неприпустимо: "кількість кібератак державні органи". З отриманих словосполучень виберіть найбільш інформативні стійки словосполучення-сутності точно з двох слів 2-грами, які входять до 3-грам. З 3-грам і 2-грам виберіть окремі найбільш інформативні одиночні слова-сутності. Вивести лише ті словосполучення, що починаються і закінчуються іменниками, а не прикметниками. Привести всі словосполучення-сутності до називного відмінку. Це - обов'язково. Вивести результати у вигляді пар. Спочатку тільки 20 основних пар типу "2-грами; 3-грами, яким вони відповідають ", потім тільки 20 пар "1-грами; 3-грами, яким вони відповідають", потім тільки 20 пар "1-грами; 2-грами, яким вони відповідають". Без повторів. Кожна пара записів в окремому рядку, наприклад, "раз два три; раз два", або "раз два; два". Не треба групувати через кому різні варіанти других понять у комі, наприклад, замість "КІБЕРАТАКИ ДЕРЖАВНІ ОРГАНИ; КІБЕРАТАКИ ДЕРЖАВНІ, ДЕРЖАВНІ ОРГАНИ" треба вивести два записи: "КІБЕРАТАКИ ДЕРЖАВНІ ОРГАНИ; КІБЕРАТАКИ ДЕРЖАВНІ" та "КІБЕРАТАКИ ДЕРЖАВНІ ОРГАНИ; ДЕРЖАВНІ ОРГАНИ". Ось текст:
За матеріалами СБУ заочні тюремні вироки отримали хакери ФСБ, які здійснили понад 5 тис. кібератак на державні органи України
Автор Марк Яшин 10.10.2024
Завдяки доказовій базі Служби безпеки та ДБР до тюремного ув'язнення заочно засуджено двох учасників хакерського угруповання ФСБ під назвою "Armageddon". За матеріалами справи, зловмисники здійснили понад 5 тис. кібератак на державні органи та об'єкти критичної інфраструктури України.

Автор Марк Яшин 10.10.2024
Завдяки доказовій базі Служби безпеки та ДБР до тюремного ув'язнення заочно засуджено двох учасників хакерського угруповання ФСБ під назвою "Armageddon". За матеріалами справи, зловмисники здійснили понад 5 тис. кібератак на державні органи та об'єкти критичної інфраструктури України.
Найбільша кількість із них припала на електронні системи Міністерства закордонних справ та Міністерство економічного розвитку нашої держави.
Як встановило розслідування, метою хакерських атак було отримати доступ до системи електронного документообігу та серверів із секретними даними урядових структур нашої держави.
За даними слідства, ключовими організаторами підривної діяльності є двоє колишніх співробітників управління СБУ в АР Крим, які у 2014 році зрадили присязі.

Тоді вони добровільно вступили до "філії" ФСБ на тимчасово окупованому півострові.
За матеріалами українських правоохоронців суд призначив обом зрадникам 15 років позбавлення волі.
Їх визнано винними за двома статтями Кримінального кодексу України:
ч. 1 ст. 111 (державна зрада);
ч. 2 ст. 361 (несанкціоновані втручання в роботу електронно-обчислювальних машин (комп'ютерів) та автоматизованих систем).
Судовий розгляд відбувся у спеціальному судовому провадженні in absentia (за відсутності обвинувачених).
Строк відбування покарання обчислюватиметься з дати фактичного затримання засуджених.
Спеціальне досудове розслідування проводили за процесуального керівництва Офісу Генерального прокурора.

DeepThink Search NEW

Рис. 2. Приклад промпту до системи DeepSeek.com

Практична реалізація інформаційної технології побудови ЛІТ реалізується за допомогою таких засобів:

Крок 1 реалізується шляхом застосування інформаційно-аналітичної системи «Кібер Агрегатор».

Кроки 2–4 реалізуються шляхом звернення до систем генеративного штучного інтелекту.

Крок 5 здійснюється людиною-експертом.

Крок 6 реалізується шляхом звернення до програми аналізу і візуалізації графів, наприклад, CSV2Graph.

У якості прикладу розглянемо текст одного документа, з якого формується ЛІТ за промптом (рис. 2), яким було здійснено звернення до системи DeepSeek.com [12].

У результаті отримано наступну відповідь від системи DeepSeek.com (рис. 3–5):

Пари "біграми; триграми, яким вони відповідають"

1. КІБЕРАТАКИ ДЕРЖАВНІ; КІБЕРАТАКИ ДЕРЖАВНІ ОРГАНИ
2. ДЕРЖАВНІ ОРГАНИ; КІБЕРАТАКИ ДЕРЖАВНІ ОРГАНИ
3. ЕЛЕКТРОННІ СИСТЕМИ; ЕЛЕКТРОННІ СИСТЕМИ МІНІСТЕРСТВА
4. МІНІСТЕРСТВА ЗАКОРДОННИХ; ЕЛЕКТРОННІ СИСТЕМИ МІНІСТЕРСТВА
5. МІНІСТЕРСТВА ЕКОНОМІЧНОГО; ЕЛЕКТРОННІ СИСТЕМИ МІНІСТЕРСТВА
6. СИСТЕМА ЕЛЕКТРОННОГО; СИСТЕМА ЕЛЕКТРОННОГО ДОКУМЕНТООБІГУ
7. СЕРВЕРИ ІЗ СЕКРЕТНИМИ; СЕРВЕРИ ІЗ СЕКРЕТНИМИ ДАНИМИ
8. ДАНИМИ УРЯДОВИХ; СЕРВЕРИ ІЗ СЕКРЕТНИМИ ДАНИМИ
9. УПРАВЛІННЯ СБУ; УПРАВЛІННЯ СБУ АР КРИМ
10. СПІВРОБІТНИКИ УПРАВЛІННЯ; СПІВРОБІТНИКИ УПРАВЛІННЯ СБУ
11. ФІЛІЯ ФСБ; ФІЛІЯ ФСБ ТИМЧАСОВО
12. ТИМЧАСОВО ОКУПОВАНОМУ; ФІЛІЯ ФСБ ТИМЧАСОВО
13. СТРОК ПОЗБАВЛЕННЯ; СТРОК ПОЗБАВЛЕННЯ ВОЛІ
14. ВОЛІ ЗАСУДЖЕНИХ; СТРОК ПОЗБАВЛЕННЯ ВОЛІ
15. СТАТТЯ КРИМІНАЛЬНОГО; СТАТТЯ КРИМІНАЛЬНОГО КОДЕКСУ
16. КОДЕКСУ УКРАЇНИ; СТАТТЯ КРИМІНАЛЬНОГО КОДЕКСУ
17. СПЕЦІАЛЬНЕ СУДОВЕ; СПЕЦІАЛЬНЕ СУДОВЕ ПРОВАДЖЕННЯ
18. ПРОВАДЖЕННЯ IN ABSENTIA; СПЕЦІАЛЬНЕ СУДОВЕ ПРОВАДЖЕННЯ
19. ДОСУДОВЕ РОЗСЛІДУВАННЯ; ДОСУДОВЕ РОЗСЛІДУВАННЯ ПРОВОДИЛИ
20. ПРОВОДИЛИ ЗА ПРОЦЕСУАЛЬНОГО; ДОСУДОВЕ РОЗСЛІДУВАННЯ ПРОВОДИЛИ

Рис. 3. Пари 2-грами та 3-грами, яким вони відповідають

Вочевидь, що завдяки здатності LLM обробляти природну мову, можна автоматично очищати текстові масиви від шуму, видаляти стоп-слова та виконувати лематизацію, що сприяє кращому виокремленню ключових термінів.

Крім того, завдяки глибоким навчальним алгоритмам, LLM здатні не лише виявляти терміни, але й формувати зв’язки між ними, виявляючи асоціативні структури, що існують у даних.

І нарешті, завдяки інтеграції LLM із системами візуалізації, такими як CSV2Graph, можна створювати інтерактивні карти ЛІТ, що дозволяє користувачам ефективно досліджувати інформацію та знаходити важливі зв’язки.

Пари "1-грами; 3-грами, яким вони відповідають"

1. КІБЕРАТАКИ; КІБЕРАТАКИ ДЕРЖАВНІ ОРГАНИ
2. ДЕРЖАВНІ; КІБЕРАТАКИ ДЕРЖАВНІ ОРГАНИ
3. ОРГАНИ; КІБЕРАТАКИ ДЕРЖАВНІ ОРГАНИ
4. ЕЛЕКТРОННІ; ЕЛЕКТРОННІ СИСТЕМИ МІНІСТЕРСТВА
5. СИСТЕМИ; ЕЛЕКТРОННІ СИСТЕМИ МІНІСТЕРСТВА
6. МІНІСТЕРСТВА; ЕЛЕКТРОННІ СИСТЕМИ МІНІСТЕРСТВА
7. СИСТЕМА; СИСТЕМА ЕЛЕКТРОННОГО ДОКУМЕНТООБІГУ
8. ДОКУМЕНТООБІГУ; СИСТЕМА ЕЛЕКТРОННОГО ДОКУМЕНТООБІГУ
9. СЕРВЕРИ; СЕРВЕРИ ІЗ СЕКРЕТНИМИ ДАНИМИ
10. ДАНИМИ; СЕРВЕРИ ІЗ СЕКРЕТНИМИ ДАНИМИ
11. УПРАВЛІННЯ; УПРАВЛІННЯ СБУ АР КРИМ
12. СПІВРОБІТНИКИ; СПІВРОБІТНИКИ УПРАВЛІННЯ СБУ
13. ФІЛІЯ; ФІЛІЯ ФСБ ТИМЧАСОВО
14. ТИМЧАСОВО; ФІЛІЯ ФСБ ТИМЧАСОВО
15. СТРОК; СТРОК ПОЗБАВЛЕННЯ ВОЛІ
16. ВОЛІ; СТРОК ПОЗБАВЛЕННЯ ВОЛІ
17. СТАТТЯ; СТАТТЯ КРИМІНАЛЬНОГО КОДЕКСУ
18. КОДЕКСУ; СТАТТЯ КРИМІНАЛЬНОГО КОДЕКСУ
19. СПЕЦІАЛЬНЕ; СПЕЦІАЛЬНЕ СУДОВЕ ПРОВАДЖЕННЯ
20. ПРОВАДЖЕННЯ; СПЕЦІАЛЬНЕ СУДОВЕ ПРОВАДЖЕННЯ

Рис. 4. Пари 1-грами, 3-грами, яким вони відповідають

№ 3 (27), 2025
ISSN 2663 - 4023

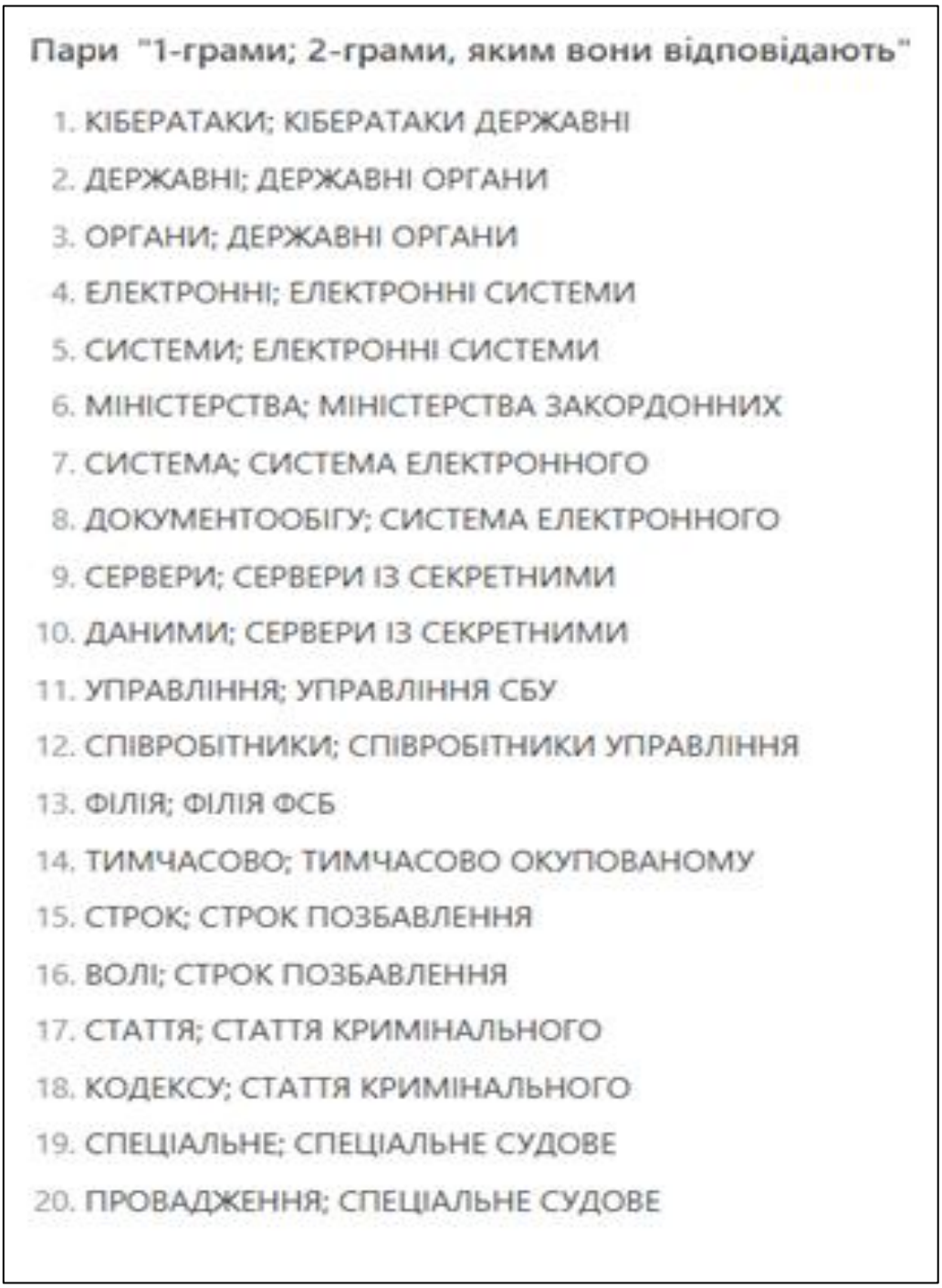

Рис. 5. 1-грами, 2-грами, яким вони відповідають

На рис. 6 представлено загальний вигляд ЛІТ розміром 20×20×20.

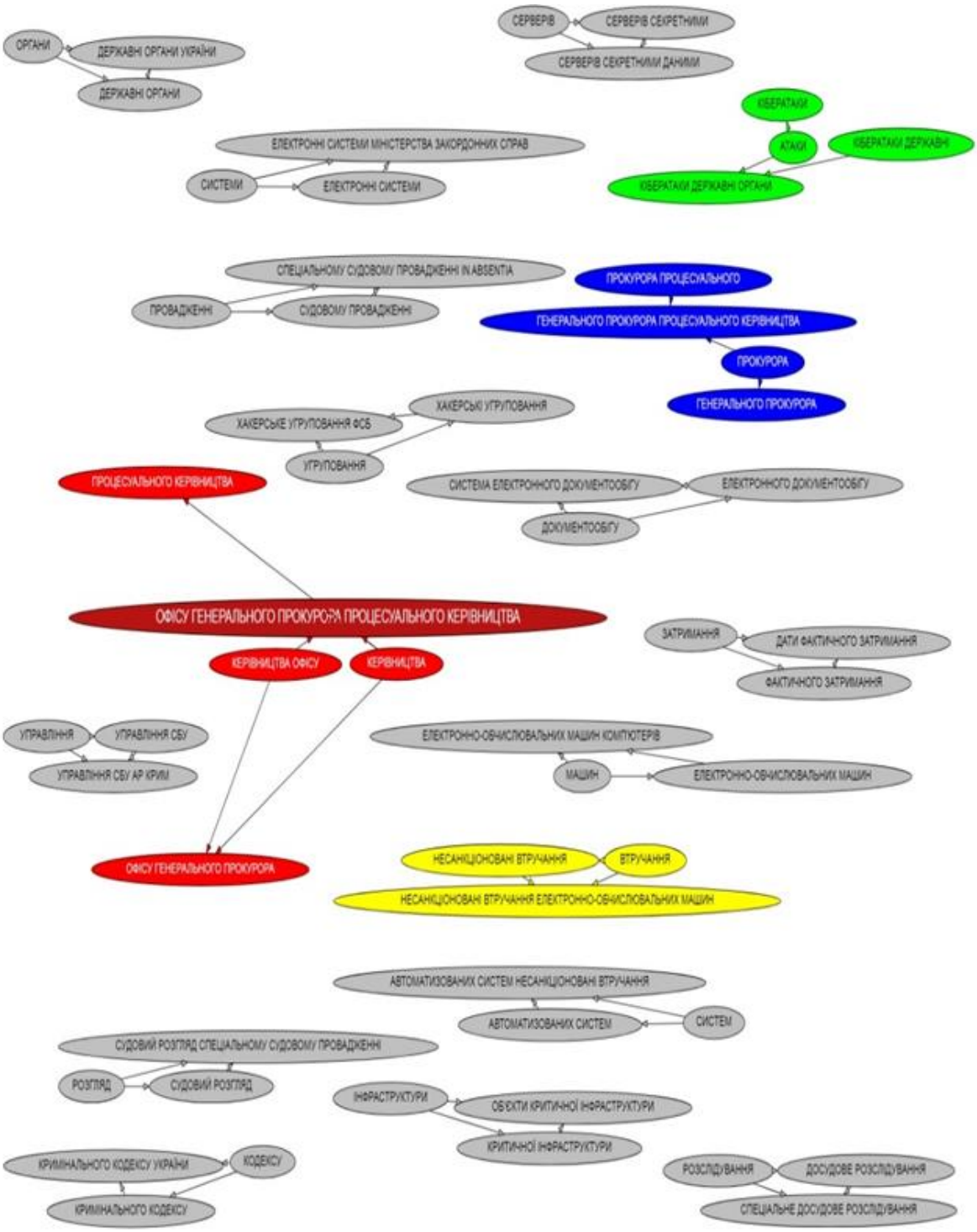

Рис. 6. Загальний вигляд ЛІТ розміром 50×50×50 (візуалізація CSV2Graph)

На рис. 7 наведено центральний фрагмент цієї мережі.

Для мережі ЛІТ за зазначеним текстовим масивом було визначено розподіл вихідних степенів вузлів, який виявився близьким до степеневого: $p(k) = Ck^{\alpha}$, тобто ця мережа є безмасштабною. Були проведені розрахунки параметрів цієї мережі. У результаті виявилось, що коефіцієнт α для неї складає близько 2,15.

Побудова ЛІТ є важливою складовою систематизації знань у сфері кібербезпеки.

У порівнянні з традиційними онтологічними підходами, метод ЛІТ забезпечує більшу гнучкість завдяки використанню LLM, що дозволяє швидше адаптуватися до нових термінів і концептів у швидко змінюваній сфері кібербезпеки. Онтології часто вимагають тривалого ручного редагування та формалізації, в той час як ЛІТ автоматизує цей процес, забезпечуючи оперативне виявлення найбільш значущих термінів.

ЛІТ пропонує більш адаптивний підхід, що дозволяє зберігати динамічність термінології та адаптуватися до змін, які відбуваються в ландшафті кіберзагроз. Це особливо важливо для професіоналів, які повинні швидко реагувати на нові виклики.

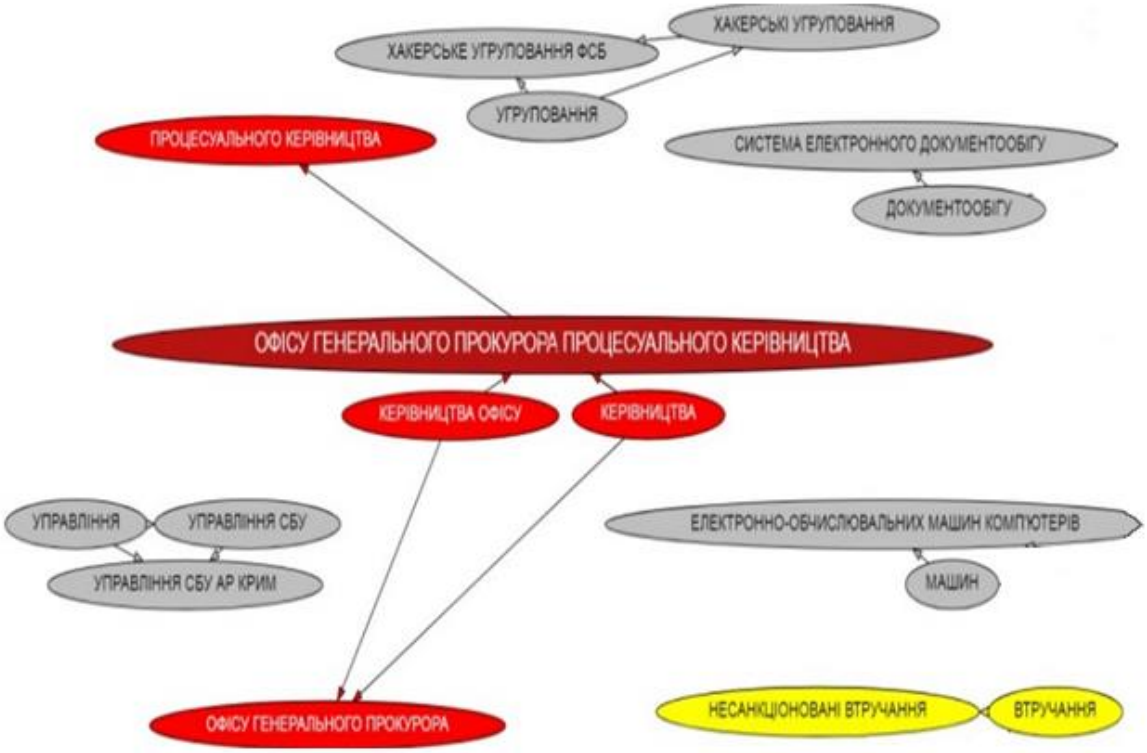

Рис. 7. Центральний фрагмент ЛІТ, побудованого за розглянутим промптом

Інші методи, такі як системи класифікації, часто мають жорсткі структури, які можуть обмежувати гнучкість при додаванні нових термінів.

У сучасній практиці, також, використовуються методи семантичного аналізу, які допомагають виявити зв'язки між термінами на основі контексту. Порівняно з цими методами, ЛІТ забезпечує більш структурований підхід до формування знань, що робить його більш ефективним для фахівців, які потребують ясної та чіткої термінології.

На практиці ЛІТ успішно застосовано для аналізу загроз у кібербезпеці. Наприклад, при оцінці нових загроз, таких як шкідливе програмне забезпечення чи фішинг, ЛІТ дозволяє швидко ідентифікувати терміни та концепти, які стосуються конкретної загрози. Це сприяє кращому розумінню загрози та формуванню адекватних заходів реагування.

Використання ЛІТ у розробці політик безпеки допомогло структурувати термінологію, що використовується в документації. Це зменшує ризик неоднозначності у формулюваннях політик, роблячи їх зрозумілішими для всіх членів команди, яка бере участь у забезпеченні кібербезпеки.

ЛІТ, також, було використано в освітніх програмах для навчання фахівців з кібербезпеки. Структуровані терміни дозволяють студентам легше засвоювати матеріал та формувати тверді знання про ключові концепти в галузі.

Застосування методики створення інтерактивної карти термінів, як це реалізовано за допомогою системи CSV2Graph, продемонструвало свою ефективність у візуалізації зв'язків між термінами. Це спрощує процес розуміння ідей і концептів, які пов'язані з кібербезпекою.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

У статті детально розглянуто важливість та методологічні аспекти побудови ЛІТ у сфері кібербезпеки. ЛІТ забезпечує гнучкість у термінології, дозволяючи адаптуватися до швидко змінюваного середовища кіберзагроз. Побудова ЛІТ є основою для подальших наукових досліджень, сприяючи глибшому розумінню проблем та розробці нових рішень.

Розроблена методика формування ЛІТ включає послідовність кроків для попередньої обробки вихідного текстового корпусу, а також вибір і сортування найбільш значущих термінів. Методика реалізована з використанням LLM, що дозволяє формулювати вимоги до ЛІТ, як задачу для цієї моделі. Це забезпечує якісне та оперативне виявлення найбільш інформативних слів і словосполучень.

Реалізована методика відображення інтерактивної карти, яка відповідає ЛІТ, за допомогою створеної авторами системи аналізу і візуалізації графів відкритого доступу CSV2Graph.

Запропоновану методику створення, аналізу та відображення ЛІТ було успішно застосовано до періодичних зведень про кібербезпеку, що створювалися на базі інформаційно-аналітичної системи «Кібер Агрегатор».

ЛІТ забезпечує чітку структуру термінів та концептів, що дозволяє систематизувати знання у галузі кібербезпеки, сприяючи кращому розумінню та використанню термінології фахівцями.

Крім того, впровадження ЛІТ спрощує процес пошуку інформації, що критично важливо для швидкого реагування на нові кіберзагрози.

Таким чином, реалізація ЛІТ є певним кроком у напрямку удосконалення управління знаннями у сфері кібербезпеки, а результати досліджень, представлені у статті, підкреслюють необхідність активного впровадження ЛІТ у практику. Це дозволить підвищити ефективність реагування на нові виклики та загрози в інформаційному середовищі.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Babayeva, G., Maennel, K., & Maennel, O. M. (2022). Building an ontology for cyber defence exercises. *2022 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, 423–432. https://doi.org/10.1109/EuroSPW55150.2022.00050
2. Adach, M., Hänninen, K., & Lundqvist, K. (2021). Structured information retrieval of security ontologies. MDH, Tech. Rep. http://www.es.mdh.se/pdf_publications/6449.pdf
3. Rackevičienė, S., & Mockienė, L. (2020). Cyber Law Terminology as a New Lexical Field in Legal Discourse. *International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique, 33(4),* 673–687. https://doi.org/10.1007/s11196-020-09690-0

4. Canito, A., Aleid, K., Praça, I., Corchado, J., & Marreiros, G. (2020). An ontology to promote interoperability between cyber-physical security systems in critical infrastructures. *2020 IEEE 6th International Conference on Computer and Communications (ICCC)*, 553–560. https://doi.org/10.1109/ICCC51575.2020.9345163
5. Satyapanich, T., Ferraro, F., & Finin, T. (2020). CASIE: Extracting cybersecurity event information from text. *Proceedings of the AAAI Conference on Artificial Intelligence, 34(5),* 8749–8757. https://doi.org/10.1609/aaai.v34i05.6401
6. Sikos, L. F. (2023). Cybersecurity knowledge graphs. *Knowledge and Information Systems, 65(12),* 3511–3531. https://doi.org/10.1007/s10115-023-01860-3
7. Ланде, Д. В. (2014). Формування мереж природних ієрархій термінів на основі аналізу текстових корпусів з правової тематики. *Правова інформатика, 1(41),* 12–17.
8. Lande, D., Snarskii, A., Yagunova, E., & Pronoza, E. (2014). Network of Natural Terms Hierarchy as a Lightweight Ontology. *Thirteenth Mexican International Conference on Artificial Intelligence (MICAI 2014)*, 16–23.
9. Lande, D. V., Snarskii, A. A., Yagunova, E. V., & Pronoza, E. V. (2013). The Use of Horizontal Visibility Graphs to Identify the Words that Define the Informational Structure of a Text. *12th Mexican International Conference on Artificial Intelligence*, 209–215.
10. Lande, D., Subach, I., & Puchkov, A. (2020). System of Analysis of Big Data from Social Media. *Information & Security: An International Journal, 47(1),* 44–61. https://doi.org/10.11610/isij.4703
11. BigSearch Space. (n.d.). https://bigsearch.space/uli.html
12. DeepSeek. (n.d.). https://www.deepseek.com/

Olexandr Puchkov
PhD in Philosophy, Professor
Head of the Institute of Special Communication and Information Protection
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
ORCID ID: 0000-0002-8585-1044
iszzi@iszzi.kpi.ua

Dmytro Lande
Doctor of Technical Sciences, Professor, Head of the Department
Educational and Scientific Physico-Technical Institute
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
ORCID ID: 0000-0003-3945-1178
dwlande@gmail.com

Ihor Subach
Doctor of Technical Science, Professor, Head of the Department
Institute of Special Communications and Information Protection
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
ORCID ID: 0000-0002-9344-713X
igor_subach@ukr.net

APPLICATION OF LARGE LANGUAGE MODELS FOR BUILDING A "FOREST OF TERM HIERARCHIES"

Abstract. One of the ways to organize and systematize knowledge is to form terminology ontologies that allow you to structure information in specific subject areas, such as cybersecurity. With the revolutionary emergence of large language models (LLMs), new opportunities are emerging to automate the process of building a "forest of term hierarchies" (FTH). Building a FTH is essential for several key aspects of cybersecurity and knowledge management, such as unifying terminology, improving communication, optimizing information retrieval, systematizing knowledge, adapting to new challenges, and supporting research and innovation. The article discusses the role of LLM in building FTH in the context of modern challenges of the information environment. Thanks to revolutionary advances in artificial intelligence, LLMs automate and optimize the processes of processing, analyzing, and structuring large amounts of text data. The article describes the key stages of FTH implementation using LLM, including text data processing, determining the discriminant power of terms, establishing links between them, and visualizing the results. A methodology for determining the associative relationships between predefined terms for building FTH is proposed. Examples of the practical implementation of the proposed methodology based on the use of the information-analytical system "Cyber Aggregator" are given. An example of forming a prompt for building a FTH for the generative artificial intelligence system DeepSeek.com is demonstrated. The technology of FTH visualization is proposed by using the program for analyzing and visualizing graphs CSV2Graph. The use of the proposed technologies makes it possible to increase the efficiency and accuracy of building terminological ontologies, which is important for adapting to the rapidly growing information flows in the modern world.

Keywords: large language models; artificial intelligence; term hierarchy forest; term ontologies; data visualization; cybersecurity.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Babayeva, G., Maennel, K., & Maennel, O. M. (2022). Building an ontology for cyber defence exercises. *2022 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, 423–432. https://doi.org/10.1109/EuroSPW55150.2022.00050

2. Adach, M., Hänninen, K., & Lundqvist, K. (2021). Structured information retrieval of security ontologies. MDH, Tech. Rep. http://www.es.mdh.se/pdf_publications/6449.pdf
3. Rackevičienė, S., & Mockienė, L. (2020). Cyber Law Terminology as a New Lexical Field in Legal Discourse. *International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique, 33(4),* 673–687. https://doi.org/10.1007/s11196-020-09690-0
4. Canito, A., Aleid, K., Praça, I., Corchado, J., & Marreiros, G. (2020). An ontology to promote interoperability between cyber-physical security systems in critical infrastructures. *2020 IEEE 6th International Conference on Computer and Communications (ICCC)*, 553–560. https://doi.org/10.1109/ICCC51575.2020.9345163
5. Satyapanich, T., Ferraro, F., & Finin, T. (2020). CASIE: Extracting cybersecurity event information from text. *Proceedings of the AAAI Conference on Artificial Intelligence, 34(5),* 8749–8757. https://doi.org/10.1609/aaai.v34i05.6401
6. Sikos, L. F. (2023). Cybersecurity knowledge graphs. *Knowledge and Information Systems, 65(12),* 3511–3531. https://doi.org/10.1007/s10115-023-01860-3
7. Lande, D. (2014). Formation of networks of natural hierarchies of terms based on the analysis of text corpora on legal topics. *Legal informatics, 1(41),* 12–17.
8. Lande, D., Snarskii, A., Yagunova, E., & Pronoza, E. (2014). Network of Natural Terms Hierarchy as a Lightweight Ontology. *Thirteenth Mexican International Conference on Artificial Intelligence (MICAI 2014)*, 16–23.
9. Lande, D. V., Snarskii, A. A., Yagunova, E. V., & Pronoza, E. V. (2013). The Use of Horizontal Visibility Graphs to Identify the Words that Define the Informational Structure of a Text. *12th Mexican International Conference on Artificial Intelligence*, 209–215.
10. Lande, D., Subach, I., & Puchkov, A. (2020). System of Analysis of Big Data from Social Media. *Information & Security: An International Journal, 47(1),* 44–61. https://doi.org/10.11610/isij.4703
11. BigSearch Space. (n.d.). https://bigsearch.space/uli.html
12. DeepSeek. (n.d.). https://www.deepseek.com/