КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (26), 2024
ISSN 2663 - 4023
DOI 10.28925/2663-4023.2024.26.663
УДК 004.942:004.912

Пучков Олександр Олександрович
к.ф.н., 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
д.т.н. професор, завідувач кафедри інформаційної безпеки
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
д.т.н. професор, завідувач Спеціальної кафедри №5
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

ЕКСТРАГУВАННЯ ОБ'ЄКТІВ КІБЕРБЕЗПЕКИ З МАСИВІВ ЕЛЕКТРОННИХ ТЕКСТОВИХ ДОКУМЕНТІВ МЕРЕЖІ ІНТЕРНЕТ ТА СОЦІАЛЬНИХ МЕРЕЖ

Анотація. Сучасний світ характеризується стрімким розвитком інформаційних технологій (ІТ) та глобальною взаємодією в кіберпросторі. Цей прогрес, незважаючи на його переваги, також призвів до виникнення нових загроз та викликів у сфері кібербезпеки. Кібервійни, які стали справжньою проблемою для держав, організацій та індивідуальних користувачів, вимагають розробки ефективних методів виявлення та аналізу об'єктів кібербезпеки. Одним з ключових аспектів у боротьбі з кіберзагрозами є можливість екстрагування фактографічних даних про об'єкти кібербезпеки з великих масивів текстової інформації. Традиційні методи аналізу тексту мають свої обмеження, особливо при роботі з великими та складними текстовими даними. У зв'язку з цим, актуальним стає застосування сучасних ІТ, які дозволяють з високою точністю та ефективністю обробляти та аналізувати текстову інформацію. У статті представлено методики екстрагування об'єктів кібербезпеки з електронних текстових документів із застосуванням регулярних виразів та виявлення об'єктів кібербезпеки на основі аналізу масивів кириличних текстів. Перша методика забезпечує виявлення фактографічних даних з текстових документів за допомогою регулярних виразів, що дозволяє точно ідентифікувати географічні назви, назви фірм та інші важливі поняття. Друга методика призначена для аналізу кириличних текстів для розпізнавання іменованих сутностей-об'єктів кібербезпеки, що спрощує процедуру екстрагування та підвищує точність отриманого результату. Кожна методика доповнює одна одну, створюючи загальну комплексну систему, яка ефективніше вирішує завдання екстрагування та аналізу об'єктів кібербезпеки порівняно з існуючими у теперішній час рішеннями. Описано алгоритми запропонованих методик, реалізація на практиці яких дозволяє з високою точністю та ефективністю обробляти та аналізувати текстову інформацію, що є важливим кроком у розробці інформаційної технології комп'ютерної розвідки з відкритих електронних джерел та соціальних мереж.

Ключові слова: кібервійна; кібербезпека; Інтернет; відкриті електронні джерела; соціальні мережі, аналіз тексту; об'єкти кібербезпеки.

ВСТУП

Сучасний світ характеризується стрімким розвитком інформаційних технологій та глобальною взаємодією в кіберпросторі. Цей прогрес, незважаючи на його переваги, також призвів до виникнення нових загроз та викликів у сфері кібербезпеки. Кібервійни, які стали справжньою проблемою для держав, організацій та індивідуальних користувачів, вимагають розробки ефективніших методів виявлення та аналізу об'єктів кібербезпеки. Одним з ключових аспектів у боротьбі з кіберзагрозами є можливість екстрагування фактографічних даних про об'єкти кібербезпеки з великих масивів текстової інформації.

Традиційні методи аналізу тексту мають свої обмеження, особливо при роботі з великими та складними текстовими даними. У зв'язку з цим, актуальним стає застосування сучасних ІТ, які дозволяють з високою точністю та ефективністю обробляти та аналізувати текстову інформацію.

Постановка проблеми. У цій статті ми представляємо методику екстрагування об'єктів кібербезпеки з масивів електронних текстових документів, розташованих в мережі Інтернет, соціальних мережах та месенджерах. Основна увага приділяється розробці підходів до виявлення фактографічних даних про об'єкти кібербезпеки з текстових документів та алгоритмів їх реалізації. Ми пропонуємо нову методику, яка включає сканування документів, порівняння з шаблонами відомих фірм, виявлення нових назв фірм, а також виявлення взаємозв'язків між поняттями та алгоритм її реалізації. Додатково до неї розглянуто методику виявлення об'єктів кібербезпеки на основі аналізу кириличних текстів.

Результати дослідження демонструють ефективність запропонованих підходів та можливість їх застосування на практиці під час вирішення завдань забезпечення кібербезпеки. Ми вважаємо, що запропоновані методики можуть стати важливим інструментом для фахівців сфери кібербезпеки, допомагаючи їм у виявленні та аналізі об'єктів кібербезпеки, а також у розробці ефективніших стратегій захисту від кіберзагроз.

Аналіз останніх досліджень і публікацій. У науковій літературі широко обговорюються методи аналізу текстових даних, включаючи використання регулярних виразів та класичних алгоритмів машинного навчання. Наприклад, у статті [1] розглянуто застосування регулярних виразів для виявлення фактографічних даних, а у роботі [2] проаналізовано ефективність класичних алгоритмів машинного навчання. Стаття [3] присвячена розробці методу автоматизованого квазіреферування електронних документів та формування множини їх ключових слів, який базується на застосуванні аналізу семантичної структури тексту та його логічної сегментації. Однак, ці методи мають свої обмеження, особливо при роботі з великими та складними текстовими даними.

Застосування таких сучасних інформаційних технологій, як технології обробки великих за обсягом даних (*Big Data*), технології аналізу текстів (*Text Mining*), а також великі мовні моделі (ВММ) та генеративний штучний інтелект (ГШІ), дозволяють подолати ці обмеження та забезпечити високу точність та ефективність обробки текстових даних.

У статті [4] детально розглянуто застосування ВММ для аналізу текстових даних, а в роботі [5] проаналізовано ефективність ГШІ у виявленні об'єктів кібербезпеки.

Крім того, у статті [6] обговорюється застосування методу *Named-entity recognition* (*NER*) для виявлення іменованих сутностей у текстах, що є важливим аспектом у вирішенні задач кібербезпеки.

У роботі [7] розглянуто використання бібліотеки *spaCy* для аналізу текстових даних, а в статті [8] проаналізовано ефективність бібліотеки *Flair* у розпізнаванні іменованих сутностей.

Робота [9] присвячена обговоренню застосування методу *Topic Modeling* для аналізу текстових даних у кібербезпеці, що дозволяє виявляти тематичні зв'язки між різними об'єктами кібербезпеки на основі аналізу текстових даних.

У статті [10] процес визначення основних об'єктів кібербезпеки та зв'язків між ними на основі аналізу змістовної складової вебпростору, розглядається в контексті єдиної інформаційної технології комп'ютерної розвідки з відкритих електронних джерел, а роботи [11] – [13] присвячені її практичній реалізації.

Проте, проведений аналіз показує, що в сучасних публікаціях відсутні ефективні методи, які об'єднували б переваги методів, що ґрунтуються на регулярних виразах, великих мовних моделях та генеративному штучному інтелекті для екстрагування об'єктів кібербезпеки з масивів текстових документів. Саме цю прогалину і намагається заповнити наша стаття, запропонувавши нові методики, які ґрунтуються на розглянутих технологіях для виявлення та аналізу об'єктів кібербезпеки.

Таким чином, зважаючи на те, що існуючи методи мають суттєві обмеження для роботи з великими за обсягом масивами текстових даних, з одного боку та сучасні досягнення у цій сфері та сфері штучного інтелекту, з іншого, стає актуальним застосування новітніх інформаційних технологій (НІТ) для розробки нових методів екстрагування об'єктів з електронних джерел, зокрема, у сфері кібербезпеки.

Мета дослідження. Загальна мета дослідження, що описується у цій статті, полягає в розробці нових методик екстрагування об'єктів кібербезпеки з використанням регулярних виразів, що містяться в масивах текстових документів, кодованих латиницею та кириличними літерами в лапках.

Для досягнення мети дослідження вирішувались такі часткові задачі:

1. Розробка методики та алгоритму виявлення фактографічних даних з текстових документів за допомогою регулярних виразів.
2. Аналіз кириличних текстів для розпізнавання іменованих сутностей-об'єктів кібербезпеки та розробка методики й алгоритму для вирішення цієї задачі.

Суть першої задачі полягає у створенні спеціально сформованих запитів інформаційно-пошуковими мовами, що включають логічні та контекстні оператори, дужки та інші елементи для виявлення географічних назв, назв фірм та інших важливих понять сфери кібербезпеки та розробці алгоритму послідовного сканування документів та порівняння з шаблонами для ефективного виявлення фактографічних даних з текстових документів.

Друга задача пов'язана з розпізнаванням іменованих сутностей на основі аналізу коротких словосполучень в інформаційному масиві латиницею або кириличними літерами в лапках та розробці методики, яка спрощує процедуру екстрагування та підвищує його точність.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Методика виявлення фактографічних даних про об'єкти кібербезпеки із застосуванням регулярних виразів

Основу запропонованої методики, зокрема, для визначення належності документа до тематичної рубрики, складають спеціально сформовані запити інформаційно-пошуковими мовами, які включають логічні та контекстні оператори, дужки та ін.

Для виявлення географічних назв об'єктів кібербезпеки, запропоновано використання таблиць, у яких, окрім шаблонів написання їх, використовуються коди країн, назви регіонів та населених пунктів.

На першому кроці на вхід методики надходить документ, який аналізується у процесі послідовного сканування.

На другому кроці текст документа порівнюється із шаблонами, що відповідають назвам відомих фірм, і якщо такі присутні, то вони містяться у спеціальній таблиці «документ–фірма».

На третьому кроці, для отримання фактографічних даних, здійснюється виявлення невідомих спочатку назв фірм на підставі, як шаблонів (регулярних виразів), так і структурних досліджень тексту. При цьому, зокрема, використовується таблиця префіксів назв фірм, що містить такі елементи, як «ТОВ», «ЗАТ», «АТ», «Компанія» та ін. відповідних запитів користувачів.

На четвертому кроці, безпосередньо за даними, представленими на ситуаційній карті, яка відбиває найактуальніші поняття (терміни, тематичні рубрики, географічні назви, імена персон, назви компаній) здійснюється виявлення взаємозв'язків понять. Таким чином, самі ситуаційні карти можуть бути вихідними даними для побудови таблиць взаємозв'язків.

Узагальнений алгоритм методики виявлення фактографічних даних із текстів із застосуванням регулярних виразів наведено на рис. 1.

Формальна постановка задачі та алгоритм методики має наступний вигляд.

Нехай **множина** $D = \{d_1, d_2, \ldots, d_N\}$ — набір текстових документів, які підлягають аналізу;

G — множина географічних назв (список шаблонів географічних назв), що включає коди країн, назви регіонів та населених пунктів;

F_{known} — **множина відомих фірм** (таблиця відомих назв фірм), зібраних з попередніх документів або баз даних;

P — **множина префіксів** (набір префіксів), що використовуються для ідентифікації назв фірм (наприклад, «ТОВ», «ЗАТ», «АТ» тощо);

R — **множина шаблонів (регулярних виразів),** які використовуються для виявлення різних типів фактографічних даних з тексту.

Необхідно: виявити фактографічні дані з *D* із застосуванням *R*.

ISSN 2663 - 4023

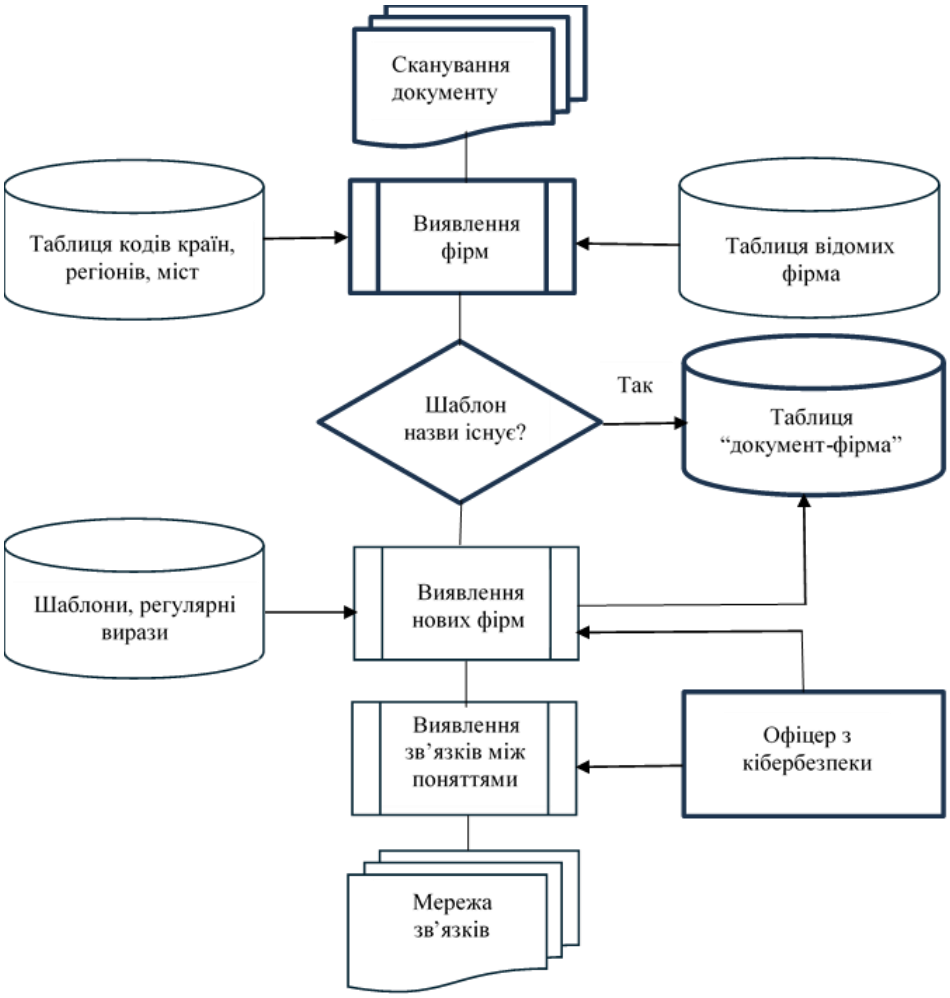

Рис. 1. Узагальнений алгоритм виявлення назв фірм із текстів документів із застосуванням регулярних виразів

Крок 1. Сканування документа. Кожен документ $d \in D$ послідовно сканується для виявлення фактографічних даних. Текст документа, представляється як послідовність символів $T_d = \{t_1, t_2, \ldots, t_M\}$, де *M* — довжина тексту.

Крок 2. Порівняння з відомими фірмами. Текст кожного документу $d \in D$ порівнюється з шаблонами назв відомих фірм з множини F_{known}. Якщо знаходяться відповідності, назва фірми додається до таблиці «документ–фірма»:

$$document\_firm(d) = \{ f \mid f \in F_{known}, match(T_d, f) \}, \quad (1)$$

де $match(T_d, f)$ — функція, що повертає істину, якщо фірма *f* зустрічається в тексті T_d документу $d \in D$.

Крок 3. Виявлення нових назв фірм. Для цього використовуються регулярні вирази *R* та структурний аналіз тексту для виявлення нових назв фірм. Для кожного префіксу $p \in P$ виконується пошук у тексті T_d документу $d \in D$:

$$F_{new}(d) = \{ f \mid f = match(T_d, p \cdot regex(R)), p \in P \}, \quad (2)$$

де $regex(R)$ — регулярний вираз для виявлення тексту, що слідує за префіксом *p*.

Нові знайдені фірми додаються до таблиці «документ–фірма»:

$$document\_firm(d) = document\_firm \cup F_{new}(d). \quad (3)$$

Крок 4. Виявлення взаємозв'язків між поняттями. На основі даних, зібраних на попередніх кроках (назви компаній, географічні назви, терміни), створюється ситуаційна карта. Мережа зв'язків формується між поняттями, які зустрічаються разом у документах або мають логічні зв'язки (наприклад, компанія *f* діє в регіоні *g*):

$$C(d) = \{(x, y) \mid x, y \in (document\_firm(d) \cup G), connected(x, y)\}, \quad (4)$$

де $connected(x, y)$ — функція, що визначає зв'язок між поняттями *x* і *y* на основі їх близькості у тексті або логічних зв'язків.

Не важко здійснити оцінку складності розглянутого алгоритму у цілому, шляхом аналізу обчислювальної складності кожного з його кроків:

- обчислювальна складність сканування (аналізу) одного документу $d \in D$ з довжиною *M* символів складає $O(M)$;
- обчислювальна складність порівняння буде $O(M \times |F_{known}|)$, де $|F_{known}|$ — кількість відомих фірм;
- обчислювальна складність виявлення нових фірм складається зі складності використання регулярних виразів і префіксів, та, відповідно залежить від кількості префіксів $|P|$ та складності застосування регулярних виразів t_{regex}: $O(M \times |P| \times t_{regex})$;
- обчислювальна складність виявлення взаємозв'язків та побудови їхньої мережі залежить від кількості знайдених понять $|X|$ і може бути оцінена, як $O(|X|^2)$.

Виходячи з наведеного, загальна обчислювальна складність запропонованого алгоритму складатиме: $O(N \times (M \times |F_{known}| + |P| \times t_{regex}) + |X|^2)$.

Наведені оцінки дозволяють стверджувати, що реалізація на практиці запропонованої методики дозволяє ефективно виявляти фактографічні дані з текстових документів.

Методика виявлення об'єктів кібербезпеки на основі аналізу масивів кириличних текстів

Зазвичай, як було наведено вище, виявлення іменованих сутностей (*Named-entity recognition, NER*) здійснюється за допомогою спеціальних програмних бібліотек (*SpaCy, Flair, FastText*), загальними недоліками яких є мала швидкість екстрагування понять і необхідність складного етапу навчання системи (відомо, що назви об'єктів кібербезпеки, назви злочинних кіберугрупувань не завжди є типовими назвами компаній або брендів).

Зокрема, бібліотека *spaCy* цікава тим, що кілька попередньо навчених моделей доступні приблизно 20 мовами [8]. Це означає, що у багатьох випадках не обов'язково навчати власну модель для отримання сутностей. Бібліотека *spaCy* вважається фреймворком «виробничого класу» тому, що вона надійна та постачається з вичерпною документацією. Ще один популярний засіб виявлення сутностей *Python* — це бібліотека *Flair* [9], що заснована на фреймворку глибокого навчання *PyTorch*. Він набуває великої

популярності, оскільки досягає вищої точності в багатьох мовах порівняно зі *spaCy*. Тому, в роботі запропонований новий підхід [10], основною ідеєю якого є розпізнавання іменних сутностей-об'єктів кібербезпеки, таких як хакерські угрупування, назви аналітичних центрів з кібербезпеки, тощо, які в повідомленнях із соціальних мереж переважно позначаються латиницею, на основі аналізу коротких словосполучень в інформаційному масиві латиницею, або кириличними літерами в лапках.

Це дозволяє значно спростити процедуру екстрагування: достатньо виявляти короткі слова або словосполучення у латинському кодуванні або у лапках. Вочевидь, технічне вирішення такої задачі не потребує великих ресурсних і часових витрат, як, наприклад, у *SpaCy*, у тому числі, спеціального машинного навчання.

При цьому для екстрагування вже відомих іменних сутностей також застосовується словник відомих іменних сутностей об'єктів кібербезпеки.

Нехай D — інформаційний масив текстів (документів), який складається з підмножин D_L (тексти латиницею) та D_C (тексти кирилицею);

S — множина іменованих сутностей, які необхідно виявити (наприклад, назви хакерських угрупувань, аналітичних центрів з кібербезпеки тощо);

T — словник відомих іменованих сутностей, зібраних з попередніх даних та мережевих джерел;

AI — функція, що представляє собою штучний інтелект, який здійснює класифікацію та визначення нових сутностей;

E — множина експертів, які можуть визначити нові сутності.

T_{new} — множина нових іменованих сутностей, визначених експертами або штучним інтелектом.

Необхідно: сформувати множину сутностей $S(d)$ у документі $d \in D$.

Суть запропонованого підходу полягає у наступному.

Спочатку, виявляються короткі словосполучення — множина коротких словосполучень $P(d)$ у документі $d \in D$, причому $P_L(d) \subseteq P(d)$ — підмножина словосполучень, написаних латиницею, а $P_S(d) \subseteq P(d)$ — підмножина словосполучень, написаних кирилицею та знаходяться в лапках.

Далі формується множина потенційних сутностей за формулою (5):

$$S'(d) = P_L(d) \cap P_C(d) \tag{5}$$

Спочатку визначаються відомі сутності. Тут для кожного $s' \in S'(d)$ необхідно перевірити, чи входить s' у словник T. Якщо $s' \in T$, додаємо його до множини виявлених сутностей $S(d)$.

Після цього визначаються нові сутності. Тут для кожного $s' \in S'(d)$, яка не входить у T, здійснюється перевірка на нову сутність. Якщо $S' \notin T$, тоді вона передається для аналізу штучному інтелекту $AI(s')$ або експерту $E(s')$. При цьому, якщо $AI(s') = positive$ або $E(s') = positive$, то s' додається до множини нових сутностей T_{new}.та оновлюється словник:

$$T = T \cup T_{new} \tag{6}$$

Остаточне визначення множини сутностей у документі *d* здійснюється наступним чином:

$$d : S(d) = S(d) \cup T_{new} \tag{7}$$

Наведемо алгоритм, який відповідає запропонованій методиці.

Крок 1. Отримати на вхід: інформаційний масив документів D, словник відомих сутностей T, функції експерта E та штучного інтелекту AI.

Крок 2. Для кожного документа $d \in D$: виявити короткі словосполучення $P(d)$, $P_L(d)$, $P_C(d)$.

Крок 3. Сформувати множину потенційних сутностей $S'(d)$ за допомогою (5).

Визначити відомі сутності $S(d)$.

Проаналізувати нові сутності за допомогою AI та E.

Крок 6. Оновити словник T (дивись формулу (6)).

Крок 7. Сформувати множину виявлених і нових сутностей $S(d)$ для кожного документу *d*, користуючись виразом (7).

Важливо зауважити, що описаний алгоритм включає кілька ключових кроків, кожен з яких має свою обчислювальну складність. Розглянемо ці кроки окремо і оцінимо загальну складність алгоритму.

Не важко помітити, що з урахуванням усіх кроків загальна обчислювальна складність має враховувати складність процесів, які на них відбуваються, а саме:

- виявлення коротких словосполучень;
- формування множини потенційних сутностей;
- визначення відомих сутностей;
- визначення нових сутностей;
- оновлення множини сутностей і словника.

Оцінимо складність кроку виявлення коротких словосполучень.

Нехай n — кількість слів у документі d. Тоді для кожного слова потрібно перевірити, чи є воно частиною короткого словосполучення. Якщо передбачити, що кожне словосполучення складається з кількох слів (наприклад, 2–3 слова), то загальна складність цього кроку буде $O(n)$.

Складність кроку формування множини потенційних сутностей *S'*(*d*). Для кожного короткого словосполучення необхідно перевірити, чи воно наведено латиницею $P_L(d)$ або кирилицею у лапках $P_C(d)$. Відповідно до цього, перевірка на належність до латиниці або кирилиці складатиме $O(1)$ часу на кожне словосполучення. Кількість коротких словосполучень вважаємо пропорційною кількості слів *n*, тому загальна складність даного кроку також складатиме $O(n)$.

Складність кроку визначення відомих сутностей *S*(*d*). Для кожної потенційної сутності $s' \in S'(d)$ перевіряємо, чи входить вона у словник T. Припускаючи, що словник реалізований у вигляді хеш-таблиці, перевірка на належність має складність $O(1)$. Тоді, загальна складність цього кроку буде $O(m)$, де m — кількість сутностей у $S'(d)$ (вважаємо пропорційною до n).

Складність кроку визначення нових сутностей T_{new}. Для кожного $s' \notin T$ застосовується штучний інтелект AI або експерт E.

Нехай t_{AI} — час для обробки однієї сутності штучним інтелектом, а t_E — час для обробки однієї сутності експертом. Тоді складність цього кроку складатиме $O(t_{AI} \times k + t_E \times k)$, де k — кількість нових сутностей.

Складність кроку оновлення множини сутностей $S(d)$ і словника T. Додавання нових сутностей до множини $S(d)$ і оновлення словника має складність $O(1)$ на кожну нову сутність. Тоді загальна складність цього кроку буде $O(k)$, де k — кількість нових сутностей.

Таким чином, відповідно до наведеного, загальна обчислювальна складність алгоритму можна оцінити як:

$$O(n)+O(n)+O(n)+O\left(t_{AI}\times k+t_E\times k\right)+O(k)=O(n)+O\left(t_{AI}\times k+t_E\times k\right) \quad (8)$$

Не важко помітити, що алгоритм має лінійну складність за кількістю слів у документі, але час виконання також залежить від складності роботи штучного інтелекту та експертів. Це робить алгоритм ефективним для обробки текстів з передбачуваною кількістю нових сутностей.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

У цій статті ми представили методики екстрагування об’єктів кібербезпеки з текстових документів, зокрема, методику виявлення фактографічних даних із застосуванням регулярних виразів, яка дозволяє ефективно виявляти фактографічні дані з текстових документів, використовуючи спеціально сформовані запити інформаційно-пошуковими мовами та шаблони географічних назв. Розроблений алгоритм методики включає послідовність кроків для послідовного сканування документів, порівняння з шаблонами відомих фірм, виявлення нових назв фірм та виявлення взаємозв’язків між поняттями. Загальна обчислювальна складність алгоритму дозволяє реалізувати його на практиці з високою ефективністю.

Методика виявлення об’єктів кібербезпеки на основі аналізу кириличних текстів дозволяє розпізнавати іменовані сутності-об’єкти кібербезпеки на основі аналізу коротких словосполучень в інформаційному масиві латиницею або кириличними літерами в лапках. Це спрощує процедуру екстрагування та зменшує необхідність складного етапу машинного навчання. Застосування словника відомих іменованих сутностей додатково підвищує точність їх виявлення.

Результати дослідження демонструють ефективність запропонованих підходів та можливість їх застосування на практиці офіцерами з кібербезпеки для вирішення задач комп’ютерної розвідки об’єктів кібербезпеки з відкритих електронних джерел та соціальних мереж і месенджерів.

Перспективними напрямками подальших досліджень є вирішення сформульованих в роботі задач шляхом використанні системи ГШІ, наприклад, *ChatGPT*, для екстрагування понять та зв’язків між ними шляхом звернення до ГШІ зі змістовними промптами та розробці алгоритму для формування мереж взаємозв’язків між суб’єктами кібербезпеки на основі екстрагованих даних, та, відповідно, визначенням об’єктів кібербезпеки-акторів кібервійни на основі даних, отриманих в результаті інформаційно-пошукових запитів до агрегаторів інформації та формування промптів до ГШІ і розробці алгоритму для побудови мережі акторів у формі графу, що враховує взаємозв’язки між ними.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Yi, F., Jiang, B., Wang, L., & Wu J. (2020). Cybersecurity Named Entity Recognition Using Multi-Modal Ensemble Learning. *IEEE Access, 8,* 63214–63224. https://doi.org/10.1109/ACCESS.2020.2984582
2. Halbouni, A., Gunawan, T. S., Habaebi, M. H., Halbouni, M., Kartiwi, M. & Ahmad, R. (2022). Machine Learning and Deep Learning Approaches for CyberSecurity: A Review. *IEEE Access, 10,* 19572–19585. https://doi.org/10.1109/ACCESS.2022.3151248
3. Субач, І., Герасимов, Б., & Сергеєв, О. (2006). Вилучення інформативних фраз із первинних електронних документів в інформаційно-пошукових системах. *УСіМ, 1,* 26–29.
4. Bayer, M., Kuehn, P., Shanehsaz, R., Reuter, C. (2024). CySecBERT: A Domain-Adapted Language Model for the Cybersecurity Domain. *ACM Transactions on Privacy and Security, 27(2(18)),* 1–20. https://doi.org/10.1145/3652594
5. Hassanin, M., & Moustafa, N. (2024). A Comprehensive Overview of Large Language Models (LLMs) for Cyber Defences: Opportunities and Directions. *arXiv preprint arXiv:2405.14487.*
6. Gao, C., et al. (2021). A review on cyber security named entity recognition. *Front. Inform. Technol. Electron. Eng. 22,* 1153–1168.
7. Hanks, C., Maiden, M., Ranade, P., Finin, T., & Joshi, A. (2022). Recognizing and extracting cybersecurity entities from text. In: *Workshop on Machine Learning for Cybersecurity, International Conference on Machine Learning.*
8. Alam, Md T., Bhusal, D., Park, Y., Rastogi, N. (2022). CyNER: A Python Library for Cybersecurity Named Entity Recognition. *arXiv preprint arXiv:2204.05754.* https://doi.org/10.48550/arXiv.2204.05754
9. Ghasiya, P., & Okamura K. (2021). Investigating Cybersecurity News Articles by Applying Topic Modeling Method. *International Conference on Information Networking (ICOIN)*, 432–438. https://doi.org/10.1109/ICOIN50884.2021.9333952
10. Lande, D., Puchkov, O., & Subach, I. (2022). Method of Detecting Cybersecurity Objects Based on OSINT Technology. In: *Selected Papers of the XXII International Scientific and Practical Conference "Information Technologies and Security" (ITS 2022), vol. 3503,* 115–124.
11. Ланде, Д. В., Субач, І. Ю., & Соболєв, А. М. (2019). *Комп'ютерна програма контент-моніторингу соціальних мереж з питань кібербезпеки* (Свідоцтво про реєстрацію авторського права на твір № 92744) КіберАгрегатор.
12. Ланде, Д. В., Субач, І. Ю., & Соболєв, А. М. (2021). *Комп'ютерна програма (картографічний сервіс) для зберігання, видачі та дослідження геоінформації* (Свідоцтво про реєстрацію авторського права на твір № 105772) ГеоАгрегатор.
13. Гулак, Г. М., Жильцов, О. Б., Киричок, Р. В., Коршун, Н. В., & Складанний, П. М. (2024). *Інформаційна та кібернетична безпека підприємства.* Підручник. Львів : Видавець Марченко Т. В.

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (26), 2024
ISSN 2663 - 4023

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, Chair 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 Special Department №5
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

EXTRACTION OF CYBERSECURITY OBJECTS FROM ARRAYS OF ELECTRONIC TEXT DOCUMENTS ON THE INTERNET AND SOCIAL NETWORKS

Abstract. The modern world is characterized by the rapid development of information technology (IT) and global interaction in cyberspace. This progress, despite its benefits, has also led to the emergence of new threats and challenges in the field of cybersecurity. Cyberwarfare, which has become a real problem for states, organizations and individual users, requires the development of effective methods for detecting and analyzing cybersecurity targets. One of the key aspects in the fight against cyber threats is the ability to extract factual data about cybersecurity objects from large amounts of textual information. Traditional text analysis methods have their limitations, especially when working with large and complex text data. In this regard, the use of modern IT, which allows processing and analyzing textual information with high accuracy and efficiency, becomes relevant. The article presents methods for extracting cybersecurity objects from electronic text documents using regular expressions and detecting cybersecurity objects based on the analysis of arrays of Cyrillic texts. The first methodology detects factual data from text documents using regular expressions, which allows for the accurate identification of geographic names, company names, and other important concepts. The second method is designed to analyze Cyrillic texts to recognize named cybersecurity entities, which simplifies the extraction procedure and increases the accuracy of the result. Each methodology complements each other, creating an overall integrated system that more effectively solves the task of extracting and analyzing cybersecurity objects compared to currently available solutions. The algorithms of the proposed methods are described, the practical implementation of which allows processing and analysing textual information with high accuracy and efficiency, which is an important step in the development of information technology for computer intelligence from open electronic sources and social networks.

Keywords: cyberwar; cybersecurity; Internet; open electronic sources; social networks, text analysis; cybersecurity objects.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Yi, F., Jiang, B., Wang, L., & Wu J. (2020). Cybersecurity Named Entity Recognition Using Multi-Modal Ensemble Learning. *IEEE Access, 8,* 63214–63224. https://doi.org/10.1109/ACCESS.2020.2984582
2. Halbouni, A., Gunawan, T. S., Habaebi, M. H., Halbouni, M., Kartiwi, M. & Ahmad, R. (2022). Machine Learning and Deep Learning Approaches for CyberSecurity: A Review. *IEEE Access, 10,* 19572–19585. https://doi.org/10.1109/ACCESS.2022.3151248
3. Subach, I., Gerasimov, B., & Sergeev, O. (2006) Extraction of informative phrases from primary electronic documents in information retrieval systems. *USiM, 1,* 26–29.
4. Bayer, M., Kuehn, P., Shanehsaz, R., Reuter, C. (2024). CySecBERT: A Domain-Adapted Language Model for the Cybersecurity Domain. *ACM Transactions on Privacy and Security, 27(2(18)),* 1–20. https://doi.org/10.1145/3652594
5. Hassanin, M., & Moustafa, N. (2024). A Comprehensive Overview of Large Language Models (LLMs) for Cyber Defences: Opportunities and Directions. *arXiv preprint arXiv:2405.14487.*
6. Gao, C., et al. (2021). A review on cyber security named entity recognition. *Front. Inform. Technol. Electron. Eng. 22,* 1153–1168.
7. Hanks, C., Maiden, M., Ranade, P., Finin, T., & Joshi, A. (2022). Recognizing and extracting cybersecurity entities from text. In: *Workshop on Machine Learning for Cybersecurity, International Conference on Machine Learning.*
8. Alam, Md T., Bhusal, D., Park, Y., Rastogi, N. (2022). CyNER: A Python Library for Cybersecurity Named Entity Recognition. *arXiv preprint arXiv:2204.05754.* https://doi.org/10.48550/arXiv.2204.05754
9. Ghasiya, P., & Okamura K. (2021). Investigating Cybersecurity News Articles by Applying Topic Modeling Method. *International Conference on Information Networking (ICOIN)*, 432–438. https://doi.org/10.1109/ICOIN50884.2021.9333952
10. Lande, D., Puchkov, O., & Subach, I. (2022). Method of Detecting Cybersecurity Objects Based on OSINT Technology. In: *Selected Papers of the XXII International Scientific and Practical Conference "Information Technologies and Security" (ITS 2022), vol. 3503,* 115–124.
11. Lande, D. V., Subach, I. Y., & Sobolev, A. M. (2019). *Computer program for content monitoring of social networks on cybersecurity issues* (Certificate of copyright registration for work No. 92744) CyberAggregator.
12. Lande, D. V., Subach, I. Y., & Sobolev A. M. (2021). *Computer program (mapping service) for storing, issuing and researching geoinformation* (Certificate of copyright registration for work No. 105772) GeoAggregator.
13. Hulak, H. M., Zhiltsov, O. B., Kyrychok, R. V., Korshun, N. V., & Skladannyi, P. M. (2024). *Information and cyber security of the enterprise.* Textbook. Lviv: Publisher Marchenko T. V.