INFORMATION TECHNOLOGY

DOI 10.20535/2411-1031.2020.8.1.217993
УДК 004[942+056.5]

ДМИТРО ЛАНДЕ,
ОЛЕКСАНДР ПУЧКОВ,
ІГОР СУБАЧ

СИСТЕМА АНАЛІЗУ ВЕЛИКИХ ОБСЯГІВ ДАНИХ З ПИТАНЬ КІБЕРБЕЗПЕКИ ІЗ СОЦІАЛЬНИХ МЕДІА

Запропоновано та обґрунтовано підходи до побудови системи моніторингу та аналізу соціальних медіа з питань кібербезпеки, які базуються на концепції обробки великих обсягів даних, складних мереж, добування знань із текстових масивів. Детально розглянуті компоненти технології Elastic Stack, інформаційно-пошукова система Sphinx, графова система управління базами даних Neo4j та система аналізу графів Gephi. Основна ідея створення системи аналізу великих обсягів даних з питань кібербезпеки із соціальних медіа – це одночасне застосування методів і засобів інформаційного пошуку, аналізу даних та агрегування інформаційних потоків. Система забезпечує реалізацію таких функцій: формування баз даних шляхом збору інформації з визначених інформаційних ресурсів; налаштування модулів автоматичного сканування і первинної обробки інформації з веб сайтів і соціальних мереж; ведення повнотекстових баз даних з інформації; виявлення дублікатів, схожих за змістом інформаційних повідомлень; повнотекстовий пошук; аналіз текстових повідомлень, визначення тональності, формування аналітичних звітів; інтеграцію з географічною інформаційною системою; аналіз та візуалізацію даних; дослідження динаміки тематичних інформаційних потоків; прогнозування розвитку подій на основі аналізу динаміки публікацій в соціальних медіа; забезпечення доступу багатьох користувачів до функціональних компонентів системи. Практичне значення отриманих результатів полягає в створенні діючого макету системи контент-моніторингу і аналізу соціальних медіа з питань кібербезпеки, який придатний до застосування як компоненти у складі систем підтримки прийняття рішень щодо інформаційної та кібербезпеки. Розглянуто інтерфейс макету системи, в якому доступні функції пошуку, аналізу та прогнозування появлення інформації в соціальних медіа. Центральне місце інтерфейсу займає дайджест із найбільш релевантних потребам користувача повідомлень. В аналітичному режимі реалізовано низку інструментів для графічного представлення аналізованих даних, які відображуються у вигляді часового ряду кількості релевантних запиту повідомлень на добу, а також перегляду головних сюжетів за темою, кластерів, згрупованих за відповідністю заздалегідь визначеним опорним словам. У системі передбачені режими формування мереж із понять, що відповідають окремим повідомленням (персон, брендів) та інформаційних джерел, які дозволяють ранжувати за рейтингом поняття та досліджувати взаємозв’язки між ними.

Ключові слова: моніторинг соціальних медіа, кібербезпека, розвідка з відкритих джерел, аналіз соціальних мереж, великі дані, Кіберагрегатор.

Постановка проблеми. На цей час, в умовах гібридної війни з розвиненою інформаційною компонентою на багатьох рівнях керування, виникає необхідність урахування інформації, що з’являється у соціальних медіа. Відомо, що інформаційні потоки інколи є компонентою інформаційних протистоянь, зміст яких спрямований на реалізацію попередньо спланованих психологічних впливів на аудиторію для досягнення заздалегідь визначених цілей. Така інформація, з одного боку, містить багато “шуму”, дезінформації, а, з

іншого боку, є найбільш оперативною. Значна кількість інформаційних ресурсів у глобальних мережах містить різні експертні оцінки, деяка частина пов'язана з реалізацією інформаційних загроз, зокрема, кібернетичних. Виходячи з цього, врахування інформації з соціальних медіа має велике значення для вирішення завдань у сфері забезпечення кібербезпеки, але до цього часу не існувало доступних бюджетних рішень проблеми цільового інформування та надання аналітичних узагальнень користувачам на основі інформації із соціальних мереж. У роботі запропоновано та обґрунтовано підходи до побудови системи моніторингу і аналізу соціальних медіа, що будується на концепції обробки великих обсягів даних (англ. "Big Data") із соціальних медіа з питань кібербезпеки та добування знань із текстових масивів (англ. "Text Mining, Information Extraction") [1].

Одним з найважливіших інструментів забезпечення кібербезпеки сьогодні виступає, так звана, розвідка за відкритими джерелами (англ. "Open Source INTelligence", OSINT) – один з напрямів розвідки, який включає в себе пошук, вибір і збір розвідувальної інформації, отриманої із загальнодоступних джерел, а також аналіз цієї інформації. Суть процесу OSINT полягає у пошуці та аналізі інформації, отриманої з відкритих джерел, збору інформації та її подальшого аналізу, а також формуванні звітів щодо об'єкту спостереження.

При створенні систем аналізу інформації із соціальних медіа необхідно вирішити проблему великих обсягів даних, що отримуються для аналізу, їх динаміки і схильності до постійних змін. Ця проблема і засоби її подолання отримали сьогодні назву "Великі дані". Підлягають вирішенню питання реалізація функцій збору, очищення, зберігання, пошуку, доступу, передачі, аналізу і візуалізації таких наборів як цілісної сутності, а не локальних фрагментів [1]. Визначальними характеристиками для великих обсягів даних є: "три V": обсяг (англ. "Volume", в сенсі величини фізичного обсягу), швидкість (англ. "Velocity", що означає в даному контексті швидкість приросту і необхідність високошвидкісної обробки і отримання результатів), різноманіття (англ. "Variety", в сенсі можливості одночасної обробки різних типів структурованих і слабкоструктурованих даних).

Сьогодні, згідно з дослідженнями агентства "Gartner", термін "Big Data" вже перекрив пік знаменитого гартнеровського "Hype Cycle". На рис. 1 наведено статистику запитів користувачів до системи Google за словосполученням "Big Data" (сервіс "Google Trends", https://trends.google.ru/). "Великі дані" – це термін, що позначає множину наборів даних настільки об'ємних і складних, що унеможливлює застосування наявних традиційних інструментів управління базами даних і застосунків для їх обробки [2].

Аналіз останніх досліджень і публікацій. Аналіз літератури, пов'язаної з темою даної статті, дозволяє сформулювати кілька важливих напрямів, пов'язаних з питаннями роботи з відкритими даними для вирішення проблем кібербезпеки, а саме: розвідку за відкритими джерелами, роботу з великими даними, аналіз соціальних мереж (англ. "Social Network Analysis", SNA), екстрагування знань з текстів, глибокий аналіз текстів (англ. "Information Extraction, Text Mining").

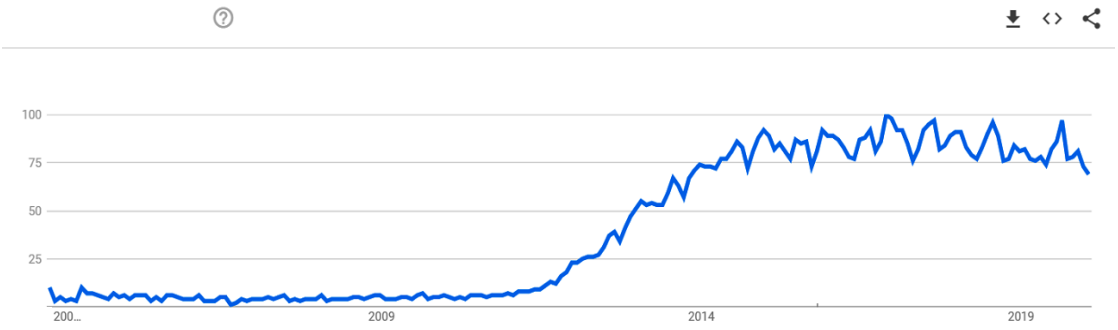

Рисунок 1 – Динаміка запитів "Big Data"

До найбільш значущих з сучасних досліджень, які належать до напряму OSINT, можна віднести роботи: [3], яка присвячена питанням автоматичного збору даних, Application Programming Interface (API) і інструментів, алгоритмів машинного навчання та застосування геоінформаційних методів; [4], в якій розглядається методологія оперативно-орієнтованих досліджень відкритих джерел для вирішення завдань забезпечення безпеки, боротьби з тероризмом і організованою злочинністю – від планування до розгортання; у [5] розглянуто завдання розвідки за відкритими джерелами в контексті боротьби з тероризмом, в тому числі, моделі, інструменти, методи і тематичні дослідження; у [6] описано тенденції використання OSINT у сфері кібербезпеки; методологія інтернет-аналітики у правових рамках розглядається у [6].

У [7] розглядається методологія аналізу великих обсягів даних, зокрема, методи математичної оптимізації та генетичні алгоритми, методи кластеризації даних, прогнозування для обробки великих обсягів даних; у [8] описано підхід до організації зберігання і обробки даних та методологія використання загальнодоступних інструментів: Hadoop, Cassandra, Cascalog, ElephantDB і Storm з Trident; робота [9] присвячена теоретичним основам та алгоритмам машинного навчання, нереляційним базам даних (НБД) типу NoSQL, обробці потокових даних, глибинному аналізу текстів, візуалізації інформації; у [10] запропоновано стратегії, архітектури та впровадження високопродуктивних рішень для сховищ даних і неструктурованих даних).

Аналіз соціальних мереж − це процес дослідження соціальних структур з використанням мереж і теорії графів. Він характеризує мережеві структури з точки зору вузлів (окремих дійових осіб, людей або речей у мережі) і зв’язків, ребер або зв’язків (відношень або взаємодій), які їх пов’язують. Прикладами соціальних структур, що зазвичай візуалізуються за допомогою спеціальних засобів, є соціальні мережі, мережі поширення мемів та інформації, мережі дружби і знайомств, мережі знань, мережі робочих відносин і співпраці, ділові мережі, соціальні мережі, мережі родинних зв’язків, інфікування. Ці мережі часто візуалізуються як соціограми, в яких вузли представлені у вигляді точок, а зв’язки представлені у вигляді ліній.

Аналізу соціальних мереж присвячені роботи: [11] – сильні і слабкі зв’язки в соціальних мережах, структурний баланс в мережах, моделювання мережевого трафіку з використанням теорії ігор, стратегічна взаємодія в мережах, інформаційні мережі; робота [12] присвячена створенню правил асоціацій на основі аналізу мереж, а також описано їх застосування в медицині, формування і дослідження мережі кодів захворювань, аналіз таких мереж за допомогою алгоритмів кластеризації; робота [13] присвячена аналізу соціальних мереж, тимчасової активності публікації сторінок, визначення шкідливих джерел інформації в мережах із застосуванням штучних нейронних мереж, машинного навчання; у [14] розглядаються методи вивчення і кластеризація мемів в соціальних мережах, подібності між текстами, кластеризація мемів, виявлення виникаючих подій, прогноз; у [15] описано мережеві аналітичні заходи, методи представлення даних у вигляді складних мереж, моделі випадкових графів, індекси центральності та їх використання в мережевому аналізі, гуманітарний аспект аналізу соціальних мереж). Глибинному аналізу інформації з соціальних мереж у даний час присвячено велика кількість робіт, серед яких можна виділити монографії [16] та [17], які присвячено вирішенню питань добування інформації з різних соціальних мереж, дослідженню прикладних програмних інтерфейсів різних мереж, аналізу текстових файлів, визначення подібності текстів, класифікація, розпізнавання образів, нейронні мережі при аналізі соціальних мереж.

Метою роботи є створення технологічних засад та інструментальних засобів аналізу вмісту соціальних мереж з питань кібербезпеки, побудова діючого макету корпоративної системи із максимальним застосуванням компонент відкритого доступу для автоматизації процесів пошуку, виявлення, моніторингу, збору, первинної обробки, аналізу, накопичення і зберігання інформації і подальшого надання її користувачам і взаємодіючим підсистемам підтримки прийняття рішень.

Виклад основного матеріалу дослідження. Розглянемо окремі функціональні компоненти, на яких має базуватися розроблена система.

Розвідка за відкритими джерелами. Відповідно до [18] OSINT базується на зборі інформації з відкритих джерел, її аналізі, підготовці і своєчасному наданні кінцевого продукту замовнику з метою вирішення розвідувальних завдань, тобто, OSINT є результатом систематизованого збору, обробки та аналізу необхідної загальнодоступної інформації. OSINT при забезпеченні кібербезпеки визначається рядом аспектів, серед яких оперативність надходження, обсяг, якість, достовірність, легкість подальшого використання та вартість отримання інформації. На процес планування та підготовки ведення OSINT впливають такі фактори [19]:

– ефективність інформаційного забезпечення, що досягається шляхом збору інформації із засобів масової інформації (ЗМІ), призначеного для користувача контенту, хештегів;

– релевантність. Доступність, глибина і масштаби публічно доступної інформації дозволяють знаходити необхідну інформацію без залучення інших спеціалізованих засобів розвідки;

– спрощення процесів збору даних. OSINT надає необхідну інформацію, виключаючи необхідність залучення зайвих технічних і людських ресурсів;

– глибина аналізу даних. Будучи частиною розвідувального процесу, OSINT дозволяє здійснювати глибокий аналіз загальнодоступної інформації з метою прийняття відповідних рішень;

– оперативність. Різке скорочення часу доступу до інформації в мережі Інтернет. Швидке отримання цінної оперативної інформації. Обстановка, що стрімко змінюється під час криз, найповніше відбивається в поточних новинах;

– обсяги. Можливість масового моніторингу певних джерел інформації з метою пошуку цільового контенту, людей і подій;

– якість. У порівнянні зі звітами спеціальних агентів, інформація з відкритих джерел позбавлена суб’єктивізму;

– достовірність;

– легкість використання. Дані OSINT можна легко передавати в будь-які зацікавлені інстанції, вони є відкритими;

– вартість. Вартість добування даних в OSINT є мінімальною.

Технології роботи з великими даними. Під час збору, аналізу відкритих даних із Інтернету виникають проблеми обробки великих обсягів, а також виникає необхідність пошуку та навігації в динамічних інформаційних потоках. Величезна кількість багатомовних динамічних інформаційних ресурсів та засилля інформаційного шуму обумовлюють складність пошуку необхідної інформації, оперативного аналізу, і звідси використання відкритих джерел в інформаційно-аналітичній роботі. Для більшості з вищеназваних проблем є актуальними питання семантичної обробки великих динамічних текстових масивів інформації. На цей час для вирішення наведених проблем застосовуються такі технологічні концепції як Big Data, Complex Networks (“Складні мережі”), Cloud Computing (“Хмарні обчислення”), Data / Text Mining (“Глибинний аналіз даних і тексту”). Усе ширше для побудови моделей предметних областей, зокрема, кібербезпеки, використовується онтологічний підхід.

За останні кілька років з’явилися різні системи для зберігання і обробки великих масивів даних. Серед них можна виділити проекти екосистеми Hadoop, нереляційні бази даних NoSQL, а також пошукові та аналітичні системи типу Elasticsearch. Hadoop і будь-яка база даних NoSQL мають свої переваги та області застосування.

Elastic Stack – екосистема компонентів, які служать для пошуку і обробки даних. Основні компоненти Elastic Stack – це Kibana, а також Logstash, Beats, X-Pack і Elasticsearch. Ядром Elastic Stack виступає пошукова система Elasticsearch, яка надає можливості для

зберігання, пошуку та обробки даних. Утиліта Kibana, яку також називають вікном в Elastic Stack, є засобом візуалізації і призначеним для користувача інтерфейсом для Elastic Stack. Компоненти Logstash і Beats дозволяють передавати дані в Elastic Stack. X-Pack надає потужний допоміжний функціонал.

Elasticsearch – швидка розподілена пошукова система повнотекстового пошуку та аналізу даних, що працює в режимі реального часу. Зазвичай Elasticsearch використовується як базовий пошуковий механізм і є основним компонентом Elastic Stack.

Elasticsearch побудована на технології Apache Lucene, тому відрізняється від традиційних рішень для реляційних баз даних або NoSQL. Нижче перераховані основні особливості використання Elasticsearch:

- неструктурованість даних, які обробляються;
- можливість пошуку;
- можливість аналізу даних;
- підтримка призначених для користувача бібліотек і REST API;
- легке управління і масштабування;
- висока швидкість роботи.

Систему Elasticsearch можна використовувати окремо без будь-яких інших компонентів Elastic Stack.

Kibana – інструмент візуалізації для Elastic Stack, який забезпечує наочне уявлення даних в Elasticsearch. У Kibana пропонується багато варіантів візуалізації інформації, таких як гістограма, карта, лінійні графіки. Kibana дозволяє в режимі реального часу створювати візуалізації і досліджувати дані в інтерактивному вигляді, а також формувати високоякісні звіти.

Elastic Stack – це гнучка платформа з розширеним набором інструментів, що дозволяє розробникам створювати власні програми завдяки великій підтримці мов програмування і REST API. Сьогодні компоненти Elastic Stack є перспективною основною побудови систем аналізу великих обсягів даних з соціальних мереж, зокрема, з проблем забезпечення кібербезпеки.

Sphinx (від SQL Phrase Index) – повнотекстова пошукова система для великих даних – Sphinx. Саме вона використовується як пошуковий механізм у побудованому на цей час діючому макеті, розробленому у рамках цієї роботи. Ця система поширюється на умовах ліцензії GNU GPL або, для версій 3.0 + без вихідних кодів. Відмінною особливістю Sphinx є висока швидкість індексації та пошуку, а також інтеграція з існуючими системами управління базами даних (СУБД, наприклад: MySQL, PostgreSQL) і API для поширених мов веб-програмування (офіційно підтримуються PHP, Python, Java; існують реалізовані спільнотою API для Perl, Ruby, .NET і C++). Система Sphinx має такі особливості:

- висока швидкість індексації (до 10-15 МБ / с на кожне процесорне ядро);
- висока швидкість повнотекстового пошуку (до 150-250 запитів в секунду на кожне процесорне ядро з 1 000 000 документів);
- велика масштабованість;
- підтримка розподіленого пошуку;
- підтримка однобайтових кодувань і UTF-8;
- підтримка морфологічного пошуку – наявні вбудовані модулі для декількох мов;
- підтримка існуючих СУБД (PostgreSQL і MySQL), а також ODBC-сумісних баз даних (MS SQL, Oracle).

Neo4j. Для обробки мереж зв'язків понять при семантичному пошуку в рамках систем аналізу великих обсягів даних із соціальних медіа, перспективною є Neo4j – графова система управління базами даних з відкритим вихідним кодом, мовою Java, з підтримкою транзакцій (Atomicity, Consistency, Isolation, Durability – ACID). На сьогодні ця система вважається найпоширенішою графовою СУБД.

Neo4j зберігає дані у власному форматі, спеціально пристосованому для подання графової інформації. Такий підхід у порівнянні із засобами реляційних СУБД дозволяє застосовувати додаткову оптимізацію. При цьому для обробки графа не потрібно завантажувати всі дані цілком в оперативну пам'ять сервера, що забезпечує обробку великих мережевих структур. Запити в Neo4j можна робити безпосередньо через Java API або мовою Gremlin, а також створеною в проекті з відкритим вихідним кодом TinkerPop мовою Cypher, яка є мовою запитів і мовою маніпулювання даними для графового сховища.

Gephi. Для забезпечення роботи з мережевими структурами на локальних робочих місцях використовується Gephi – пакет програмного забезпечення з відкритим кодом для мережевого аналізу та візуалізації. Gephi (https://gephi.org/) – це в даний час найпопулярніша програма візуалізації і аналізу мереж та графів ("Мережевих графів"). Пакет Gephi забезпечує швидку компоновку, ефективну фільтрацію та інтерактивне дослідження даних, а також є одним з кращих варіантів для візуалізації великомасштабних мереж на персональних комп'ютерах. Gephi дозволяє оброблювати графові структури досить великих обсягів (до 1 млн. вузлів) на персональному комп'ютері завдяки застосуванню ефективних алгоритмів. Пакет Gephi – це мультиплатформне програмне забезпечення, яке розповсюджується з відкритим кодом згідно з ліцензіями CDDL 1.0 і GNU General Public License v3. За адресою https://gephi.org/ доступні версії вихідних кодів для Mac OS X, Windows і Linux. Розробники Gephi позиціонують цю програму, як "Photoshop, але для даних".

Програма включає в себе множину різних алгоритмів компонування графів і дозволяє настроювати в них кольори, розміри і мітки. Пакет Gephi є інтерактивним програмним забезпеченням і надає засоби для виявлення спільнот, а також надає можливість розрахунку найкоротших шляхів або відносної відстані від будь-якого вузла до даного. Плагіни від Gephi дозволяють розширювати її функціональність і додавати нові алгоритми, макети та інструменти вимірювань. Gephi має багатопотокову схему обробки даних, і таким чином, дозволяє виконувати кілька видів аналізу одночасно.

Пакет Gephi постачається з ефективними алгоритмами компонування, такими як Yifan-Hu, Force-directed та надає можливість: завантаження даних мереж в форматах GEXF, GDF, GML, GraphML, Pajek (NET), GraphViz (DOT), CSV, UCINET (DL), Tulip (TPL), Netdraw (VNA) і таблиць Excel; експортування даних мереж в форматах JSON, CSV, Pajek (NET), GUESS (GDF), Gephi (GEFX), GML та GraphML. Завдяки цьому пакет Gephi може взаємодіяти з іншими системами аналізу і візуалізації графів.

Функціональні можливості системи. Актуальним підходом вирішення проблеми створення описаної системи, є одночасне застосування методів і засобів інформаційного пошуку, аналізу даних і агрегування інформаційних потоків. У межах роботи побудовано та досліджено діючий макет системи моніторингу і аналізу соціальних медіа, автоматичної обробки динаміки і повних текстів із соціальних мереж за заданий період часу, які пов'язані із тематикою "Кібербезпека".

Збирання інформації здійснюється у режимі пошуку в соціальних медіа (веб сайтах, соціальних мережах, месенджерах, блогах). Запит (ключова фраза для пошуку у відповідній соціальній мережі, якщо це можливо, інакше – аккаунт) зчитується програмою із спеціальних конфігураційних таблиць. Далі здійснюється пошук і виведення записів, що відповідають запитам. Після цього унікальні записи зберігаються у БД серверу.

Аналіз існуючих підходів до агрегації тематичних новин спонукав до необхідності створення комплексу інструментальних засобів контент-моніторингу соціальних мереж з вибраних питань, зокрема, забезпечення кібербезпеки [20].

Макет включає сучасні засоби персоналізації, надання доступу до баз даних в режимі онлайн, у тому числі з мобільних пристроїв, для чого широко застосовуються можливості форматів Really Simple Syndication (RSS). Обґрунтовано вибір "готових" програмних компонентів, описані засоби власної розробки (сканери соціальних мереж, засоби формування динамічних RSS-каналів), наведені результати їх інтеграції у єдиний програмно-апаратний комплекс.

Система аналізу великих обсягів даних із соціальних медіа повинна забезпечити реалізацію наведених нижче функцій:

1) формування баз даних шляхом підключення до мережі Інтернет та збору за певними критеріями та акаунтами інформації, наведеної у національних кодуваннях з визначених інформаційних ресурсів (на першому етапі, надалі перелік буде збільшено): веб сайтів; блогів (Twitter, Livejournal); соціальних мереж (Facebook, Instagram, Reddit, Medium); відеохостингів (YouTube, RuTube); наукових спільнот (Academia.edu, ArXiv.org); месенджерів (Telegram);

2) налаштування адміністратором системи модулів автоматичного сканування і первинної обробки веб сайтів і соціальних мереж. За необхідності створення службових акаунтів, через які буде організований доступ до визначених соціальних мереж;

3) ведення ретроспективних повнотекстових баз даних з інформації, що збирається із Інтернету; створення, ротація баз даних і забезпечення формування внутрішніх словникових наборів даних, наведених різними мовами (індексування повідомлень) в цих базах даних з використанням універсальної системи кодування – Unicode Transformation Format-8 (UTF-8);

4) виявлення дублікатів, схожих за змістом інформаційних повідомлень (зокрема, різними мовами), групування дублікатів та близьких за змістом інформаційних повідомлень для видачі пошуковою системою;

5) реалізація повнотекстового пошуку із застосуванням запитів, наведених різними мовами;

6) первинний аналіз текстових повідомлень, що зберігаються в базах даних системи: автоматичне виявлення іменованих сутностей (осіб, назви компаній, брендів, географічні назви), визначення тональності, виявлення опорних слів за статистичними алгоритмами в інформаційних матеріалах, наведених різними мовами;

7) формування аналітичних звітів, а також інформаційних портретів і сюжетних ланцюжків, що ґрунтуються на використанні опорних слів, наведених різними мовами, тематична рубрикація документів;

8) інтеграція з геоінформаційною системою;

9) аналіз та візуалізація даних; візуалізація статистичних даних: за визначеними джерелами, кількістю завантажених повідомлень за період часу; графіки (гістограми) розподілу кількості інформаційних повідомлень, із зазначенням розподілу кількісних показників за джерелами, типами джерел, датою;

10) застосування вейвлет-аналізу для дослідження тематичних інформаційних потоків. Технологія використання вейвлетів дозволяє виявляти одиничні та нерегулярні “сплески”, різкі зміни значень кількісних показників у різні періоди часу, зокрема, обсягів тематичних публікацій у соціальних мережах. При цьому можуть виявитися ситуації виникнення циклів, а також ситуації, коли за періодами регулярної динаміки настають хаотичні коливання. Визначено, що динаміку інформаційних операцій найточніше відображують такі відомі вейвлети, як “Мексиканський капелюх” та вейвлет Морле [21];

11) прогнозування розвитку подій на основі аналізу динаміки публікацій в соціальних медіа. Для дослідження часових рядів обсягів повідомлень у тематичних інформаційних потоках, сьогодні все ширше використовується теорія фракталів та методи нелінійного аналізу [22]. Проте часові ряди, породжувані тематичними інформаційними потоками, також мають фрактальні властивості та їх можна розглядати як стохастичні фрактали. Такий підхід розширює сферу застосування теорії фракталів на інформаційні потоки, динаміка яких описується засобами теорії випадкових процесів. Крім того, для прогнозування можуть застосовуватися хвильові методи, що використовуються на цей час також для аналізу фінансових ринків [23];

12) забезпечення доступу багатьох користувачів до функціональних компонентів системи, розмежування доступу щодо перегляду робіт, що виконуються користувачами.

Основу апаратної платформи систем аналізу великих обсягів даних із соціальних медіа складають такі сервери:

– інформаційний проксі-сервер (орендований віртуальний сервер, що забезпечує прихований збір інформації, розташований на зовнішньому дата центрі. При розвитку системи таких серверів може бути декілька. Цей сервер, з одного боку, призначений для надійного обслуговування користувачів корпоративних мереж, а з іншого – може забезпечувати обмін даними з аналогічними зовнішніми проксі-серверами);

– сервер добування даних (основний сервер збору даних із Інтернет-ресурсів. Може добувати дані за визначеними адміністратором сценаріями безпосередньо з Інтернет-ресурсів, або з інформаційних проксі-серверів);

– сервер первинної аналітики (на сервері здійснюється первинна аналітична обробка інформації, а також інформаційний пошук. За допомогою сервера підтримуються бази даних ретроспективної інформації. Первинна аналітична обробка інформації охоплює: добування понять; геоінформаційну підтримку; визначення тональності повідомлень; формування зведень; аналіз динаміки повідомлень; прогнозування; аналіз масиву джерел інформації);

– фронтенд-сервер (веб-сервер, з якого забезпечується доступ кінцевих користувачів через веб браузери, RSS-агрегатори або через API програмних застосувань до ресурсів системи).

Інтерфейс макету системи аналізу великих обсягів даних із соціальних медіа. Реалізований макет системи аналізу великих обсягів даних із соціальних медіа “Кіберагрегатор” [24], [25] надає користувачеві веб інтерфейс, з якого йому доступні функції пошуку та аналізу інформації в соціальних медіа (див. рис. 2).

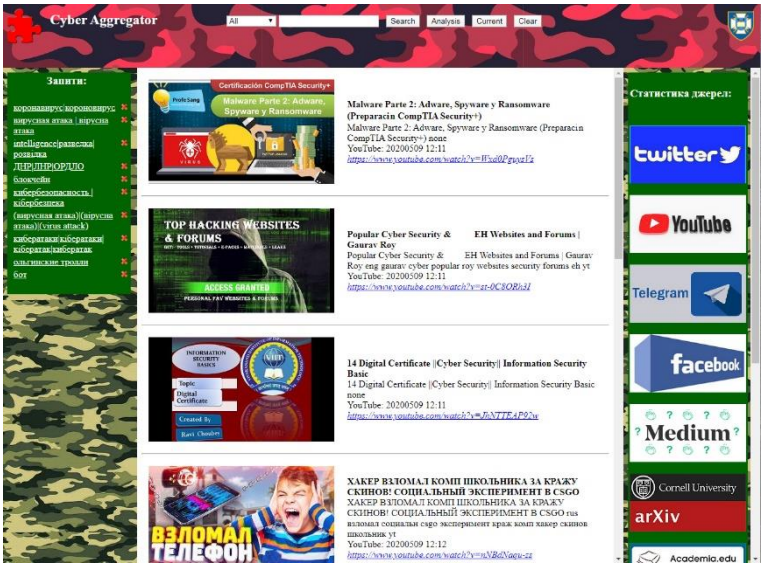

Рисунок 2 – Інтерфейс користувача системи “Кіберагрегатор”

Користувачу системи надаються можливості пошуку (як у ретроспективній базі даних повідомлень із соціальних медіа (“Search”), так й у поточній інформації (“Current”), а також аналізу даних (“Analysis”). Центральне місце інтерфейсу займає дайджест із найбільш релевантних потребам користувача повідомлень. В окремому блоці “Запити” відображаються збережені користувачем запити. Статистична інформація щодо поповнення бази даних системи з окремих соціальних медіа доступна у блоці “Статистика джерел”.

У результаті пошуку за запитом (див. рис. 3) користувачу надається перелік заголовків відповідних запиту (релевантних) повідомлень із гіперпосиланнями на повні тексти цих повідомлень в системі, а також на ці повідомлення в соціальних медіа.

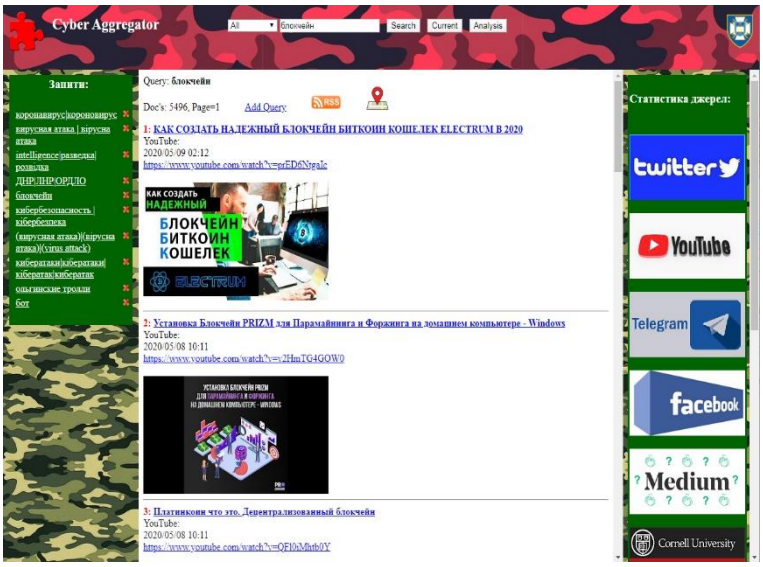

Рисунок 3 – Фрагмент інтерфейсу користувача в пошуковому режимі (результати пошуку за запитом “Блокчейн”)

Якщо на запит система видає відповідні інформаційним потребам користувача документи, то його можна зберегти для подальшого застосування (“Add Query”). Можливе подальше виведення знайдених повідомлень у форматі RSS (із подальшим завантаженням цих результатів в, так звані, RSS-агрегатори на постійній основі), а також виведення результатів пошуку із деталізацією на географічній карті, яка масштабується як в автоматичному режимі, так і шляхом налаштувань (див. рис. 4).

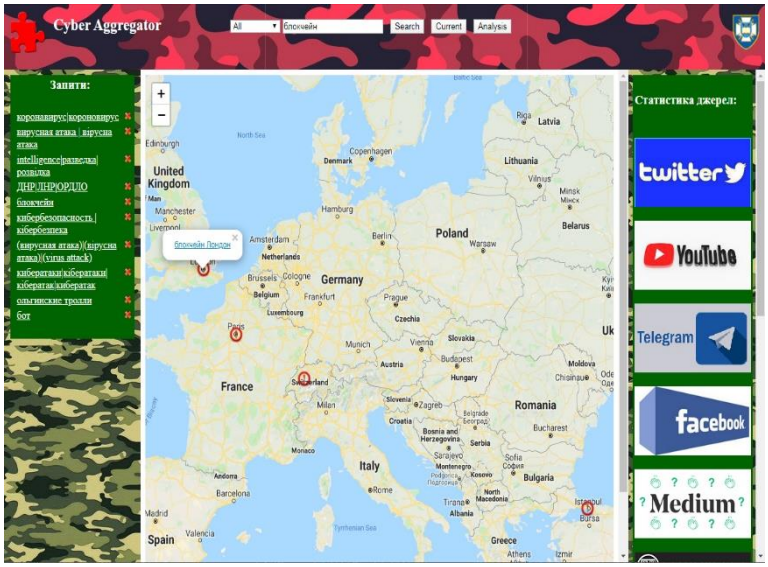

Рисунок 4 – Фрагмент інтерфейсу із застосуванням геоінформаційної системи

В аналітичному режимі (“Analysis”) користувачеві доступна низка інструментів, перший з яких – це графік (“Graph”), що відповідає часовому ряду кількості релевантних запиту повідомлень на добу (див. рис. 5). Користувачеві, також, є доступною можливість перегляду головних сюжетів (“Digest”) за темою (див. рис. 6), кластерів, згрупованих за відповідністю заздалегідь визначених опорних слів.

У системі передбачені режими формування мереж із понять, що відповідають окремим повідомленням (персон, брендів), інформаційних джерел (див. рис. 7). Ці режими дозволяють ранжувати за рейтингом поняття та досліджувати взаємозв’язки між ними.

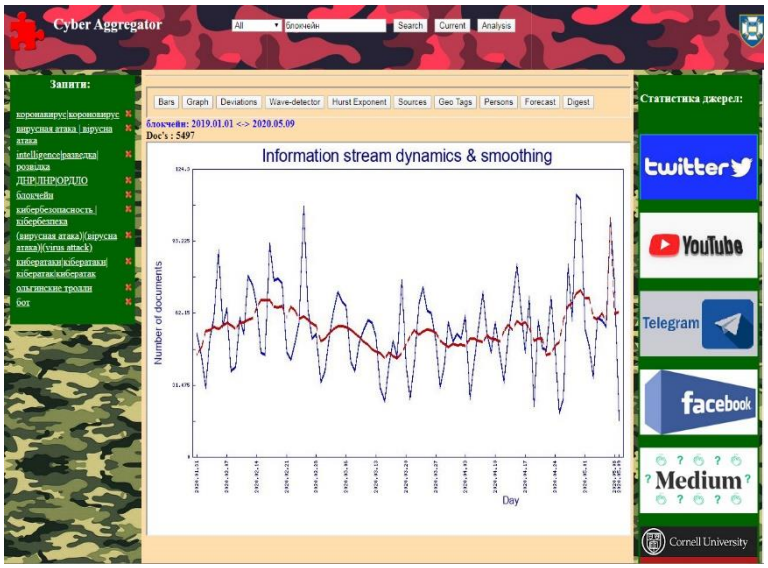

Рисунок 5 – Динаміка тематичних повідомлень за запитом “Блокчейн”

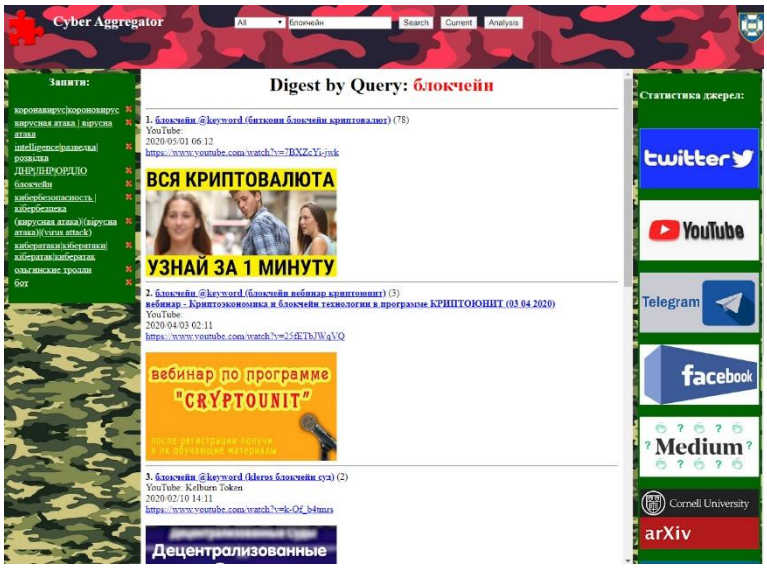

Рисунок 6 – Фрагмент тематичного дайджесту за запитом “Блокчейн”

У режимі “Аналітика” передбачено можливість прогнозування (“Forecast”) методом, запропонованим Д. Сорнетте [26], [27], який заснований на аналізі закономірності руху ринкових цін на товарних і фондових ринках перед крахом. У роботі показано, що перед крахом ціна має степеневе зростання, ускладнене логоперіодичними коливаннями, які сходяться до нескінченності в критичній точці, де ймовірність краху досягає максимальної величини. Відповідна степенева модель, яка враховує лінійні логоперіодичні коливання, має наступний вигляд:

$$F(t) = A + B(t_c - t)^m \left[1 + C\cos\left(\omega \log\left(\frac{t_c - t}{T} \right) + \varphi \right) \right].$$

де $F(t)$ – степенева модель з врахуванням логоперіодичних коливань;

t_c – критичний час (час кризи);

$A,\ B,\ \omega,\ \varphi$ – коефіцієнти моделі, що визначаються за допомогою процедури підбору.

За допомогою використання моделі Сорнетте (клавіша “Forecast”, див. рис. 8.) на основі даних моніторингу можна отримати прогнозні значення логарифму кількості відповідних публікацій.

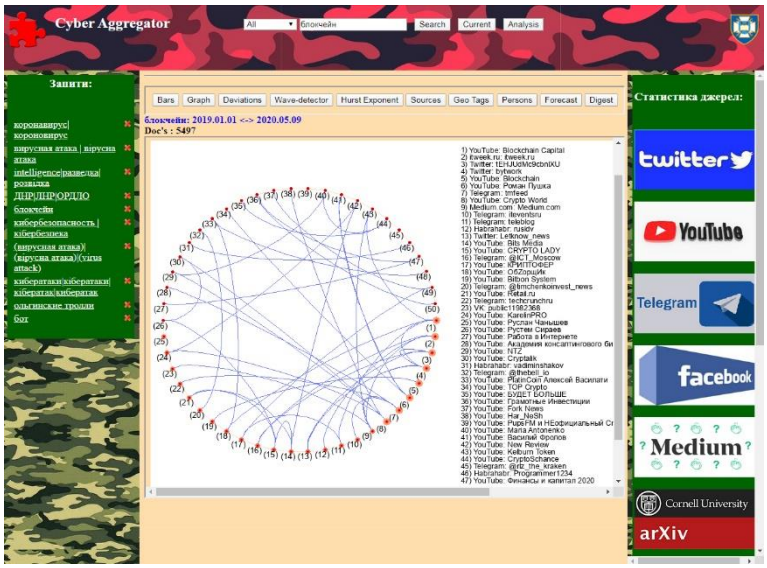

Рисунок 7 – Мережа взаємозв'язку джерел інформації, на яких публікувалися повідомлення за запитом "Блокчейн"

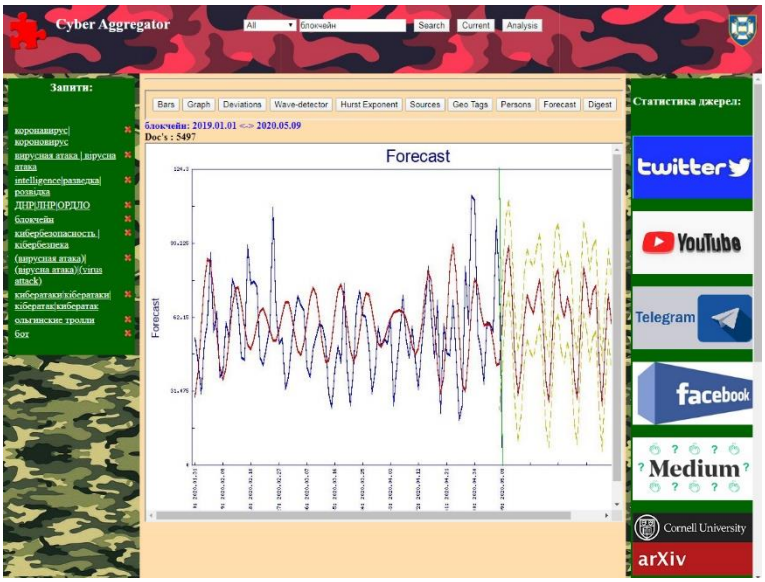

Рисунок 8 – Прогнозна лінія за алгоритмом Сорнетте для часового ряду, що відповідає запиту "Блокчейн"

Висновки. На цей час в усьому світі створюються і впроваджуються програмні та технологічні рішення для систем аналізу великих обсягів даних із соціальних медіа, які використовуються при автоматизації заходів з інформаційної та кібербезпеки, підвищують ефективність аналізу великих масивів інформації, дозволяють розпізнавати природні і змодельовані інформаційні потоки, виявляти зв'язки учасників інформаційних процесів і задіяні інформаційні ресурси. Розпізнавання інформаційних впливів, атак, операцій дозволяє виявляти та організовувати протидію, відстежувати інформаційну та кібербезпеку систем, уразливість критично важливих інфраструктур.

Запропоновано та обґрунтовано інформаційні технології створення системи контент-моніторингу соціальних мереж за визначеною проблематикою, вибору релевантної інформації із соціальних мереж, реалізацію інформаційно-пошукового механізму їх уточнення користувачами, збереження запитів як RSS-каналів, ведення персональних баз даних у середовищі клієнтських застосувань.

Практичне значення отриманих результатів полягає в створенні діючого макету системи контент-моніторингу і аналізу соціальних медіа з питань кібербезпеки, готового до застосування як компоненти систем підтримки прийняття рішень щодо інформаційної і кібербезпеки.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Д. В. Ланде, І. Ю. Субач, та Ю. Є. Бояринова, *Основи теорії і практики інтелектуального аналізу даних у сфері кібербезпеки*, Київ, Україна: ІСЗЗІ КПІ ім. Ігоря Сікорського, 2018.

[2] D. Boyd, and K. Crawford, "Critical questions for Big Data", *Journal Information, Communication & Society*, vol. 15, iss. 5, pp. 662-679, 2012, doi:10.1080/1369118X.2012.678878.

[3] R. Layton, and P. A. Watters, *Automating open source intelligence: algorithms for OSINT*: Elsevier, Syngress, 2016, doi: 10.1016/C2014-0-02170-3.

[4] B. Akhgar, P. S. Bayerl, and F. Sampson, *Open Source Intelligence Investigation. From Strategy to Implementation*: Springer International Publishing AG, 2016.

[5] N. Memon, and R. Reda Alhajj, *Counterterrorism and Open Source Intelligence*, Wien, Austria: Springer-Verlag, 2011.

[6] E. J. Appel, *Cybervetting. Internet Searches for Vetting, Investigations, and Open-Source Intelligence*: Taylor & Francis Group, LLC, 2015.

[7] J. W. Foreman, *Data Smart. Using Data Science to Transform Information into Insight*: Wiley, 2013.

[8] N. Marz, and J Warren, *Big Data: Principles and best practices of scalable realtime data systems*: Manning, 2012.

[9] Д. Силен, А. Мейсман, та М. Али, *Основы Data Science и Big Data. Python и наука о данных*, Санкт-Петербург, Россия: Питер, 2017.

[10] K. Krishnan, *Data Warehousing in the Age of Big Data*: Elsevier Inc, 2013.

[11] D. Easley, and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*: Cambridge University Press, 2010.

[12] G. Ragozini, and M. P. Vitale, *Challenges In Social Network Research: Methods And Applications: Lecture Notes In Social Network*: Springer, 2020.

[13] M. Kaya, J. Kawash, S. Khoury, and M. Y. Day, *Social Network Based Big Data Analysis and Applications*: Springer International Publishing, 2018.

[14] M. Kaya, Ö. Erdogan, and J. Rokne, *From Social Data Mining and Analysis to Prediction and Community Detection*: Springer International Publishing, 2017.

[15] K. A. Zweig, *Network Analysis Literacy: A Practical Approach to the Analysis of Networks*, Wien, Austria: Springer-Verlag, 2016.

[16] M. A. Russell, and M. Klassen, *Mining the Social Web Data Mining Facebook Twitter LinkedIn Instagram*: O'Reilly Media, 2019.

[17] M. A. Russell, *21 Recipes for Mining Twitter*: O'Reilly Media, 2011.

[18] ATP 2-22.9, Army Techniques Publication, no. 2-22.9 (FMI 2-22.9). Headquarters Department of the Army Washington, DC, 10 July 2012.

[19] D. Lande, and E. Shnurko-Tabakova, "OSINT as a part of cyber defense system", *Theoretical and Applied Cybersecurity*, no. 1, pp. 103-108, 2019, doi: 10.20535/tacs.2664-29132019.1.169091.

[20] Д. В. Ланде, "Аналіз інформаційних потоків у глобальних комп'ютерних мережах", *Вісник НАН України*, № 3, с. 46-54, 2017, doi: 10.15407/visn2017.03.045.

[21] A. Dodonov, D. Lande, V. Tsyganok, O. Andriichuk, S. Kadenko, and A. Graivoronskaya, *Information Operations Recognition. From Nonlinear Analysis to Decision-Making*: Lambert Academic Publishing, 2019.

[22] П. А. Кисельов, та Д. В. Ланде, “Розробка програмних засобів аналізу та прогнозування інформаційних операцій”, на *Науково-практичній конференції курсантів (студентів), аспірантів, докторантів та молодих вчених “Актуальні питання застосування спеціальних інформаційно-телекомунікаційних систем”*, Київ, 2019, с. 180.

[23] О. Г. Додонов, Д. В. Ланде, О. В. Нестеренко, та Б. О. Березін, “Підхід до прогнозування дієвості державного управління з використанням технологій OSINT”, *на XIX Международной научно-практической конференции ИТБ-2019*, Киев, 2019, с. 230-233.

[24] Д. В. Ланде, І. Ю. Субач, та А. М. Соболєв, “Комп’ютерна програма контент-моніторингу соціальних мереж з питань кібербезпеки “Кіберагрегатор” (“Кіберагрегатор”), Свідоцтво про реєстрацію авторського права на твір № 91831, лип. 31, 2019.

[25] Д. В. Ланде, Н. А. Кальян, та О. Т. Матіїшин, “Система агрегації соціальних медіа з питань кібербезпеки”, на *XVII Всеукраїнській науково-практичній конференції студентів, аспірантів та молодих вчених “Теоретичні і прикладні проблеми фізики, математики та інформатики”*, Київ, 2019, с. 10-11.

[26] Д. Сорнетте, *Как предсказывать крахи финансовых рынков. Критические события в сложных финансовых системах*, Litres, 2017.

[27] О. В. Уренцов, “Проверка возможности предсказания кризисов на финансовом рынке с помощью метода Д. Сорнетте”, *Труды Института системного анализа Российской академии наук*, № 40, с. 174-191, 2008.

Стаття надійшла до редакції 03.03.2020.

REFERENCE

[1] D. V. Lande, I. Yu. Subach, and Yu. Ye. Boyarinova, *Fundamentals of the theory and practice of data mining in the field of cyber security*, Kyiv: Institute of special communication and information protection of National technical university of Ukraine “Igor Sikorsky Kyiv polytechnic institute”, 2018.

[2] D. Boyd, and K. Crawford, “Critical questions for Big Data”, *Journal Information, Communication & Society*, vol. 15, iss. 5, pp. 662-679, 2012, doi:10.1080/1369118X.2012. 678878.

[3] R. Layton, and P. A. Watters, *Automating open source intelligence: algorithms for OSINT*: Elsevier, Syngress, 2016, doi: 10.1016/C2014-0-02170-3.

[4] B. Akhgar, P. S. Bayerl, and F. Sampson, *Open Source Intelligence Investigation. From Strategy to Implementation*: Springer International Publishing AG, 2016.

[5] N. Memon, and R. Reda Alhajj, *Counterterrorism and Open Source Intelligence*, Wien, Austria: Springer-Verlag, 2011.

[6] E. J. Appel, *Cybervetting. Internet Searches for Vetting, Investigations, and Open-Source Intelligence*: Taylor & Francis Group, LLC, 2015.

[7] J. W. Foreman, *Data Smart. Using Data Science to Transform Information into Insight*: Wiley, 2013.

[8] N. Marz, and J Warren, *Big Data: Principles and best practices of scalable realtime data systems*: Manning, 2012.

[9] D. Cielen, A. Meysman, and M. Ali, *Introducing Data Science. Big Data, Machine Learning, and More, Using Python Tools*: Manning Publications Co., 2016.

[10] K. Krishnan, *Data Warehousing in the Age of Big Data*: Elsevier Inc, 2013.

[11] D. Easley, and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*: Cambridge University Press, 2010.

[12] G. Ragozini, and M. P. Vitale, *Challenges In Social Network Research: Methods And Applications: Lecture Notes In Social Network*: Springer, 2020.

[13] M. Kaya, J. Kawash, S. Khoury, and M. Y. Day, *Social Network Based Big Data Analysis and Applications*: Springer International Publishing, 2018.

[14] M. Kaya, Ö. Erdogan, and J. Rokne, *From Social Data Mining and Analysis to Prediction and Community Detection*: Springer International Publishing, 2017.
[15] K. A. Zweig, *Network Analysis Literacy: A Practical Approach to the Analysis of Networks*, Wien, Austria: Springer-Verlag, 2016.
[16] M. A. Russell, and M. Klassen, *Mining the Social Web Data Mining Facebook Twitter LinkedIn Instagram*: O'Reilly Media, 2019.
[17] M. A. Russell, *21 Recipes for Mining Twitter*: O'Reilly Media, 2011.
[18] ATP 2-22.9, Army Techniques Publication, no. 2-22.9 (FMI 2-22.9). Headquarters Department of the Army Washington, DC, 10 July 2012.
[19] D. Lande, and E. Shnurko-Tabakova, "OSINT as a part of cyber defense system", *Theoretical and Applied Cybersecurity*, no. 1, pp. 103-108, 2019, doi: 10.20535/tacs.2664-29132019.1. 169091.
[20] D. Lande, "*Information Streams Analysis in the Global Computer Networks*", *Visnyk NAS of Ukraine*, no. 3, pp. 46-54, 2017, doi: 10.15407/visn2017.03.045.
[21] A. Dodonov, D. Lande, V. Tsyganok, O. Andriichuk, S. Kadenko, and A. Graivoronskaya, *Information Operations Recognition. From Nonlinear Analysis to Decision-Making*: Lambert Academic Publishing, 2019.
[22] P. Kisel'ov, and D. Lande, "Development of software for analysis and forecasting of information operations", *in Proc. of the scientific-practical conference of cadets (students), graduate students, doctoral students and young scientists "Topical issues of special information and telecommunications systems*", Kyiv, 2019, pp. 180.
[23] O. Dodonov, D. Lande, O. Nesterenko, and B. Berezin, "Approach to forecasting the effectiveness of public administration using OSINT technologies", *in Proc. of the XIX International Scientific and Practical Conference ITS-2019*, Kyiv, 2019. pp. 230-233.
[24] D. Lande, I. Subach, and A. Sobolyev, "Computer program "Computer program of social networks content monitoring on cybersecurity "CyberAggregator" ("CyberAggregator")", Ukraine, Certificate of registration of copyright to the work № 91831, July 31, 2019.
[25] D. Lande, N. Kalyan, and O. Matiishin, "Social media aggregation system on cybersecurity", *in Proc. of the XVII All-Ukrainian scientific-practical conference of students, graduate students and young scientists "Theoretical and applied problems of physics, mathematics and computer science"*, Kyiv, 2019, pp. 10-11.
[26] D. Sornette, *How to predict the collapse of financial markets. Critical events in complex financial systems*, Litres, 2017.
[27] O. V. Urentsov, "Testing the possibility of predicting crises in the financial market using the method of D. Sornette", *in Proc. of the Institute of System Analysis of the Russian Academy of Sciences*, 2008, no. 40, pp. 174-191.

DMYTRO LANDE,
OLEKSANDR PUCHKOV,
IHOR SUBACH

SYSTEM FOR ANALYSING OF BIG DATA ON CYBERSECURITY ISSUES FROM SOCIAL MEDIA

The paper proposes and substantiates approaches to building a corporate system for monitoring and analyzing social media on cybersecurity issues, which are based on the concept Big Data, Data/Text Mining, Information Extraction, Complex Networks, and Cloud Computing. The components of Elastic Stack technology, Sphinx information retrieval system, Graph Data Base Management System Neo4j, and Gephi graph analysis system are examined in detail. The main idea of a system for analyzing large amounts of data on cybersecurity issues from social media is the simultaneous application of methods and means of information retrieval, data analysis, and aggregation of information flows. The system should ensure the implementation of the following

functions: the formation of databases by collecting information from certain information resources; settings for automatic scanning and primary processing of information from websites and social networks; maintaining full-text information databases; identifying duplicates similar in content to informational messages; full-text search; analysis of text messages, determination of tonality, the formation of analytical reports; integration with geographic information system; data analysis and visualization; study of the dynamics of thematic information flows; predicting developments based on the analysis of the dynamics of the publication in social media; providing access for many users to the functional components of the system. The practical significance of the results is to create a working layout of the content monitoring and analysis system of social media on cybersecurity issues, ready to be used as a component in information and cybersecurity decision support systems. The interface of the system layout is considered, in which the functions of search, analysis, and forecasting of information appearance in social media are available. Central to the interface is a digest of the most relevant user needs. In the analytical mode, a number of tools are implemented for graphical presentation of the analyzed data, which are displayed as a time series of the number of relevant queries per day, as well as viewing the main topics, clusters grouped by predefined reference words. The system provides modes for forming networks of concepts that correspond to individual messages (persons, brands) and information sources that allow you to rank the concepts and explore the relationships between them.

Keywords: social media monitoring, cybersecurity, open-source intelligence, social media analysis, Big Data, CyberAggregator.

Ланде Дмитро Володимирович, доктор технічних наук, професор, завідувач відділом спеціалізованих засобів моделювання, Інститут проблем реєстрації інформації Національної академії наук України, Київ, Україна.

ORCID: 0000-0003-3945-1178.

E-mail: dwlande@gmail.com.

Пучков Олександр Олександрович, кандидат філософських наук, професор, начальник, Інститут спеціального зв'язку та захисту інформації Національного технічного університету України "Київський політехнічний інститут імені Ігоря Сікорського", Київ, Україна.

ORCID: 0000-0002-8585-1044.

E-mail: iszzi@iszzi.kpi.ua.

Субач Ігор Юрійович, доктор технічних наук, доцент, завідувач кафедри кібербезпеки і застосування інформаційних систем і технологій, Інститут спеціального зв'язку та захисту інформації Національного технічного університету України "Київський політехнічний інститут імені Ігоря Сікорського", Київ, Україна.

ORCID: 0000-0002-9344-713X.

E-mail: igor_subach@ukr.net.

Lande Dmytro, doctor of technical science, professor, head at the specialized modeling tools department, Institute for information recording of National academy of science of Ukraine, Kyiv, Ukraine.

Puchkov Oleksandr, candidate of philosophy science, professor, head, Institute of special communication and information protection National technical university of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine.

Subach Ihor, doctor of technical science, associate professor, head at the cybersecurity and application of information systems and technologies academic department, Institute of special communication and information protection National technical university of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine.