Агрегація інформації з питань кібербезпеки як основа навчального курсу «Оброблення надвеликих масивів даних»

Aggregation of information on cybersecurity as a basis for the training course "Processing of large data sets"

Dmytro Lande
Інститут проблем реєстрації інформації Національної академії наук України, Київ, http://orcid.org/0000-0003-3945-1178
Oleksandr Puchkov
Інститут спеціального зв'язку та захисту інформації Національного технічного університету України "Київський політехнічний інститут імені Ігоря Сікорського", Київ, http://orcid.org/0000-0002-8585-1044
Ihor Subach
Інститут спеціального зв'язку та захисту інформації Національного технічного університету України "Київський політехнічний інститут імені Ігоря Сікорського", Київ, http://orcid.org/0000-0002-9344-713X
DOI:
https://doi.org/10.20535/2411-1031.2020.8.1.217993

Анотація

На цей час в галузі кібернетичної безпеки все більшу роль грає поняття «великих даних» (Big Data). Звичайно, кількість даних, яку необхідно враховувати в галузі кібербезпеки постійно зростає, разом з цим зростають й обсяги інформаційного шуму, інколи деструктивного характеру. Фахівці, що займаються обробкою, агрегацією великих обсягів даних, вирішенням проблем, обумовлених їх зростанням, динамікою, варіативністю на цей час називають «вченими з обробки даних» (Data Scientists), відповідно, наука – Data Science.

В роботі обґрунтовано і представлено основні положення навчального курсу «Оброблення надвеликих масивів даних» як введення в спеціальність Data Science в сфері кібербезпеки, на основі вивчення теоретичних основ цієї спеціальності і практичного застосування відповідних інформаційних технологій агрегації великих обсягів даних.

В рамках курсу «Оброблення надвеликих масивів даних» розглядаються базові, найпоширеніші сьогодні технології і інструменти в області кібербезпеки, перелік яких дозволяє отримати досить цілісне уявлення про те, що використовують сьогодні фахівці в області Data Science і інструменти, якими необхідно володіти, щоб вести проекти з використанням великих даних.

Предметом навчальної дисципліни є фундаментальні положення про концепцію "великих даних"; відповідні моделі даних; архітектурні концепції створення інформаційних систем для "великих даних"; аналітика "великих даних", а також питання практичного застосування результатів обробки "великих даних". Дисципліна включає два розділи: «Великі дані: теоретичні засади», і «Технологічні застосування для великих даних» і десять тем в рамках цих розділів, які розглянемо детально.

Як екосистема, полігон для проведення практичних занять в рамках курсу розглядається макет на основі системи «КіберАгрегатор», який створено і постійно удосконалюється в рамках даного курсу.

Система «КіберАгрегатор» складається з трьох основних частин, це сервер для збору та первинної обробки інформації, сервер пошуку інформації (пошукова система) та інтерфейсний сервер, з якого послуга надається користувачам та іншим системам через API. Система базується на таких технологічних компонентах, як інформаційно-пошукова система Elasticsearch, утілити Kibana, графових систем керування базами даних Neo4j, засобів візуалізації результатів на основі JavaScript (D3.js) і модулі сканування мережевої інформації. Система забезпечує реалізацію таких функцій, як формування баз даних з визначених інформаційних ресурсів; 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 тощо.

В результаті проходження курсу студенти отримують знання і навички, необхідні для ефективної обробки великих обсягів даних із соціальних мереж,створення систем моніторингу мережевої інформації з питань кібербезпеки, відбору релевантної інформації із соціальних мереж, впровадження пошукової системи, проведенні аналітичних досліджень, прогнозування.

Ключові слова: Big Data, Соціальні мережі, Навчальний курс, Інформаційно-пошукові системи, Агрегація даних, Data Science

Currently, the concept of Big Data is playing an increasingly important role in the field of cybersecurity. Of course, the amount of data that needs to be considered in the field of cybersecurity is constantly growing. At the same time, the volume of information noise, sometimes of a destructive nature, is also increasing. Specialists involved in processing, aggregating of Big Data, solving problems caused by their growth, dynamics, variability are now called Data Scientists, respectively, science - Data Science.

The work substantiates and presents the main principles of the training course "**Processing of large data sets**" as an introduction to the Data Science specialty in the field of cybersecurity, based on the study of the theoretical foundations of this specialty and the practical application of the relevant information technologies for aggregating large amounts of data.

As part of the course "**Processing of large data sets**", the basic, most common technologies and tools in the field of cybersecurity are considered. The list considered in the work allows you to get a fairly holistic idea of what data scientists use today and the tools that you need to have in order to run projects using Big Data.

The subject of the academic discipline is the fundamental provisions on the concept of Big Data; appropriate data models; architectural concepts for creating information systems for Big data; analytics of Big Data, as well as questions of practical application of the results of processing "big data". The discipline includes two sections: "Big Data: Theoretical Foundations", and "Technological Applications for Big Data" and ten topics within these sections, we will consider in detail.

As an ecosystem, a testing ground for practical exercises within the course, a model based on the CyberAgregator system is considered, which was created and is constantly being improved within the framework of this course.

The CyberAgregator system consists of three main parts: a server for collecting and primary processing of information, an information retrieval server (search engine) and an interface server from which the service is provided to users and other systems via API. The system is based on such technological components as the information retrieval system Elasticsearch with tools Kibana, graphs of Neo4j database management systems, visualization tools based on JavaScript (D3.js) and network information scanning modules. The system provides the formation of databases for certain information resources; maintaining full-text information databases; identification of duplicates with similar content of information messages; full-text search; analysis of text messages, determination of sentiment, formation of analytical reports; integration with a geographic information system; data analysis and visualization; study of the dynamics of thematic information flows; forecasting the development of events based on the analysis of the dynamics of publications and the like.

As a result of completing the course, students acquire the knowledge and skills necessary for the effective processing of Big Data from social networks, the creation of monitoring systems for network information on cybersecurity issues, the selection of relevant information from social networks, the implementation of a search system, analytical research, and forecasting.

Keywords: Big Data, Social Networks, Training Course, Information Retrieval Systems, Data Aggregation, Data Science

Біографії авторів

Dmytro Lande, Інститут проблем реєстрації інформації Національної академії наук України, Київ, доктор технічних наук, професор, завідувач відділом спеціалізованих засобів моделювання

Oleksandr Puchkov, Інститут спеціального зв'язку та захисту інформації Національного технічного університету України "Київський політехнічний інститут імені Ігоря Сікорського", Київ, кандидат філософських наук, професор, начальник

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

Постановка проблеми

На цей час в галузі кібернетичної безпеки все більшу роль грає поняття «великих даних» (Big Data) [1]. Звичайно, кількість даних, яку необхідно враховувати в галузі кібербезпеки постійно зростає, разом з цим зростають й обсяги інформаційного шуму, інколи деструктивного характеру. Фахівці, що займаються обробкою, агрегацією великих обсягів даних, вирішенням проблем, обумовлених їх зростанням, динамікою, варіативністю на цей час називають «вченими з обробки даних» (Data Scientists), відповідно, наука – Data Science.

Великі дані – це термін, що позначає множину наборів даних настільки об'ємних і складних, що унеможливлює застосування наявних традиційних інструментів управління базами даних і додатків для їх обробки. Проблему представляють збір, очищення, зберігання, пошук, доступ, передача, аналіз і візуалізація таких наборів як цілісної сутності, а не локальних фрагментів. В якості визначальних характеристик для великих даних відзначають «три V»: обсяг (Volume, в сенсі величини фізичного обсягу), швидкість (Velocity, що означає в даному контексті швидкість приросту і необхідність високошвидкісної обробки і отримання результатів), різноманіття (Variety, в сенсі можливості одночасної обробки різних типів структурованих і напівструктурованих даних).

Виходячи з цього, актуальним на цей час є підготовка фахівців із спеціальності Data Science, що мають знання і навички з роботи технологіям Big Data, формування відповідних наукових курсів, що охоплюють необхідні теоретичні відомості і вміння роботи із відповідними інформаційними технологіями.

Метою цієї роботи є обґрунтування і представлення основних положень навчального курсу «Оброблення надвеликих масивів даних» як введення в спеціальність Data Science в сфері кібербезпеки, на основі вивчення теоретичних основ цієї спеціальності і практичного застосування відповідних інформаційних технологій агрегації великих обсягів даних.

Аналіз останніх досліджень і публікацій

Термін Big Data вперше з'явився в редакційній статті Кліффорда Лінча, редактора журналу Nature, 3 вересня 2008 року, який присвятив цілий спеціальний випуск того журналів темі "що можуть значити для сучасної науки набори великих даних".

На цей час існує ряд публікацій, пов'язаних із роллю Data Science, необхідністю підготовки фахівців з цього напрямку, це, насамперед, робота Біла Френкса [1], Деві Сілена,

Арно Мейсмана [3], Додонова О.Г. [4] (|розділ щодо аналітики великих даних), разом з цим комплексного навчального курсу, пов’язаного з реальною практичною базою, яку побудовано на сучасному вільному програмному забезпеченні, досі не існує.

Розвиток напрямку Big Data пов’язано з розвитком соціальних мереж, незважаючи на те, що великі обсяги даних притаманні таким також галузям, як телекомунікаційна, енергетична, транспортна тощо. І однією з перших інформаційних технологій в сфері безпеки і оборони, пов’язаних з цією проблемою є OSINT (Open Source Intelligence, розвідка у відкритих джерелах). В роботі [2] обґрунтовано, що OSINT є складовою частиною кібербезпеки.

У відповідності із [1, 3] можна виділити такі основні функціональні операції над великими даними:

- агрегація (консолідація) даних;
- класифікація, кластеризація;
- машинне навчання;
- візуалізація.

Виклад основного матеріалу

В рамках курсу «Оброблення надвелікіх масівів даних» розглядаються базові, найпоширеніші сьогодні технології і інструменти в області кібербезпеки, розглянутий нижче перелік яких не вичерпує всіх апробованих технологій, проте він дозволяє отримати досить цілісне уявлення про те "що" користуються сьогодні фахівці в області Data Science і інструменти, якими необхідно володіти, щоб вести проекти з використанням великих даних.

Предметом навчальної дисципліни є фундаментальні положення про концепцію “великих даних”; відповідні моделі даних; архітектурні концепції створення інформаційних систем для “великих даних”; аналітика “великих даних”, а також питання практичного застосування результатів обробки “великих даних”.

В рамках навчальної дисципліни розглядаються:

- причини виникнення нового напрямку великих даних та проблем і можливостей, пов'язаних з появою великих даних;
- можливості технологій аналізу надвеликих масивів даних для вирішення проблем підприємств, організацій чи бізнесу, а також можливостей застосування наукових методів, у т.ч. методів інтелектуального аналізу даних, до великих даних;
- особливості архітектурних рішень при створенні та розгортанні систем обробки великих масивів даних, а також вибору технології зберігання й обробки великих даних, використання сучасних високопродуктивних систем зберігання й обробки великих даних;
- основні технології і інструменти роботи з великими даними: Hadoop, HDFS, MapReduce, Elastic Stack, Elasticsearch, Kibana, Neo4j;
- компоненти програмного забезпечення, необхідні для роботи в розподілених інформаційних системах обробки надвеликих даних.

Структура дисципліни, основні розділи

Дисципліна включає два розділи: «Великі дані: теоретичні засади», і «Технологічні застосування для великих даних» і десять тем в рамках цих розділів, які розглянемо детально.

Тему 1 присвячено введенню в Big Data, розглядаються концептуальні положення, питання агрегації (концептуальні), кластеризації і класифікації, машинного навчання, візуалізації. В межах цієї теми розглядаються визначення та термінологія великих даних, роль великих даних техніці, науці, економіці та суспільному житті. Вивчаються такі характеристики великих даних, як об’єм, швидкість, різноманітність. Також розглядаються

можливі джерела великих даних (дані соціальних мереж; персональні дані; сенсорні дані; дані моніторингових систем; дані транзакцій; адміністративні дані).

Відповідно, тему 2 присвячено Data Science – сучасній науці про дані. Розглядаються основні поняття, сфери застосування, питання машинного навчання при обробленні надвеликих масивів даних. Також розглядаються елементи інформаційних технологій, що включають реалізацію алгоритмів машинного навчання для великих даних. Для управління даними на цей час застосовуютьс так звані NoSQL [5] системи керування базами даних (СКБД). Розглядаються особливості розробки інформаційних систем на базі NoSQL-рішень на прикладах СКБД СУБД MongoDB, CouchDB та Redis.

Теми 3 і 4 присвячено методам класифікації і кластерного аналізу великих даних. Дається Визначення класифікації та кластерного аналізу як базових методик інтелектуального аналізу великих даних. Розглядається співвідношення класифікації та кластеризації. При цьому класифікація – це машинне навчання з учителем (Supervised Machine Learning), кластерний аналіз – машинне навчання без вчителя (Unsupervised Machine Learning). Вивчається математична формалізація процесів класифікації і кластерного аналізу як задач оптимізації. Розглядаються такі алгоритми класифікації, як метод *k*-найближчих сусідів, лінійний класифікатор, ДНФ-метод, метод опорних векторів (SVM), і кластерного аналізу, такі як, методи k-means, ієрархічного агрегування (HAC), матричного латентного семантичного індексування (LSI).

В рамках п’ятої теми розглядаються такі фундаментальні поняття Data Science , як нейронні мережі і машинне навчання (machine learning, ML) як засіб інтелектуального аналізу великих масивів даних. способи машинного навчання.

Шосту тему присвячено концепції складних мереж (Complex Networks), які розглядаються як спеціальний вид великих даних. Вивчаються основні поняття концепції складних мереж, окремі параметри і властивості складних мереж, серед яких розподіл ступенів вузлів складних мереж, кластерність, модулярність тощо.

Наступні чотири теми складають другий розділ (змістовий модуль), який присвячено технологічним платформам великих даних, серед яких технологія Apache Hadoop [6], призначена для організації розподіленого оброблення великих об’ємів даних, MapReduce – технологія розподіленого паралельного оброблення великих масивів даних з використанням великого числа обчислювальних кластерів. Основними засобом агрегації великих обсягів неструктурованих даних, їх пошуку, обробки візуалізації , в рамках цього курсу розглядається екосистема компонентів Elastic Stack [7, 8], що служать для пошуку і обробки даних. Детально розглядаються основні компоненти цього стеку, а саме, Kibana [8], Logstash, Beats, X-Pack і Elasticsearch. Elasticsearch – це інформаційно-пошукова система, ядро Elastic Stack, яка дозволяє здійснювати обробку неструктурованих даних, інформаційний пошук, аналіз даних, забезпечує підтримку призначених для користувача бібліотек і REST API; легке управління і масштабування. Утиліта Kibana – це вікно в Elastic Stack, засіб візуалізації, що реалізує такі види відображення даних із Elasticsearch , як гістограми, карти, лінійні графіки, часові ряди.

10 тему присвячено засобам аналізу великих мереж, графовим СКБД. Розглядаються можливості двох основних систем – програми забезпечення для аналізу та візуалізації графів Gephi [11, 12] і графова система керування базами даних Neo4j [13]. Серед особливостей програми Gephi вивчаються інтерфейс користувача, можливості компонування графів, фільтрація, дослідження даних, візуалізація, підтримка графічних форматів даних. Графова СУБД Neo4j забезпечує збереження і обробку мережевих даних великих обсягів, містить декларативну мову запитів до графів Cypher .

Практичні заняття за цим курсом надають навички застосування практично усіх наведених інформаційних технологій на основі макету (система «КіберАгрегатор»), який створено і постійно удосконалюється в рамках даного курсу.

Макет. Практична реалізація

В ході практичних занять вирішуються завдання створення інтелектуальної інформаційно-пошукової системи на основі технологій стеку Elastic, її заповнення даними, що збираються із веб-сторінок і соціальних мереж, агрегації цих даних, створення засобів аналітичної обробки цих даних, виявлення трендів, прогнозування тощо. Також передбачається автоматизоване формування моделей предметних областей, їх візуалізація.

Як і більшість подібних систем для агрегування інформації із соціальних мереж, система «КіберАгрегатор» [14, 15] складається з трьох основних частин (серверів) (Рис. 1), це сервер для збору та первинної обробки інформації, сервер пошуку інформації (пошукова система) та інтерфейсний сервер, з якого послуга надається користувачам та іншим системам через API.

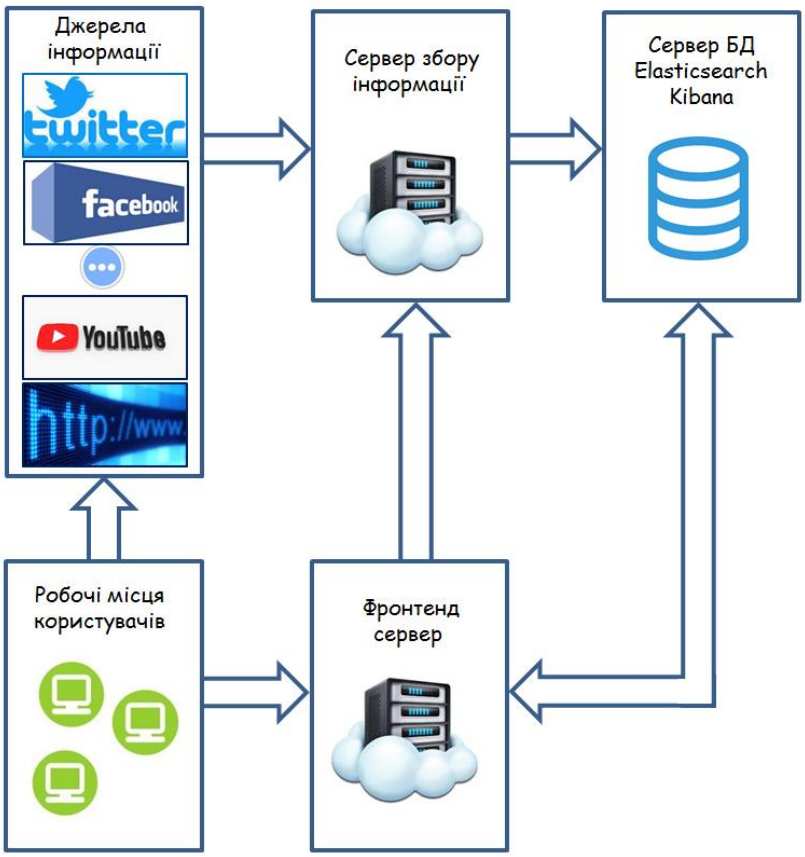

Рисунок 1: Схема потоків інформації в системі «КіберАгрегатор»

Основою апаратної платформи систем для аналізу великих даних із соціальних мереж досі є сервери:

- інформаційний проксі-сервер (орендований віртуальний сервер, що забезпечує анонімний збір інформації, розташований у зовнішньому центрі обробки даних. З розвитком системи таких серверів може бути кілька. Цей сервер, з одного боку, призначений для надання надійних послуг користувачі корпоративних мереж, а з іншого боку, він може забезпечити обмін даними з подібними зовнішніми проксі);
- сервер збору даних (сервер для збору даних з Інтернет-ресурсів. Він може витягувати дані за сценаріями, визначеними адміністратором безпосередньо з Інтернет-ресурсів, або через інформаційні проксі-сервери);

- сервер аналітики (сервер здійснює аналітичну обробку інформації та пошук інформації. За допомогою сервера підтримуються бази даних історичної інформації. Аналітична обробка інформації включає: вилучення понять; підтримка геоінформації; визначення тональності повідомлень; формування інформації; аналіз динаміка повідомлень; прогнозування; аналіз масиву джерел інформації тощо);
- інтерфейсний сервер (веб-сервер, з якого кінцеві користувачі можуть отримати доступ через веб-браузери, RSS-агрегатори або через API додатків до системних ресурсів).

Функціонування систем агрегування інформації із соціальних мереж включає такі етапи (Рис. 2):

1) пошук повідомлень із соціальних мереж, які мають відношення до загальної широкої теми - формування інформаційного потоку з тематичних повідомлень;

2) визначення мови окремих повідомлень, які завантажуються із соціальних мереж;

3) витяги з інформаційних повідомлень, таких понять, як ключові слова, особи, компанії, географічні назви тощо;

4) аналіз тональності окремих повідомлень;

5) форматування даних, перетворення в стандартні формати (XML, JSON);

6) завантаження отриманого потоку в повнотекстові бази даних.

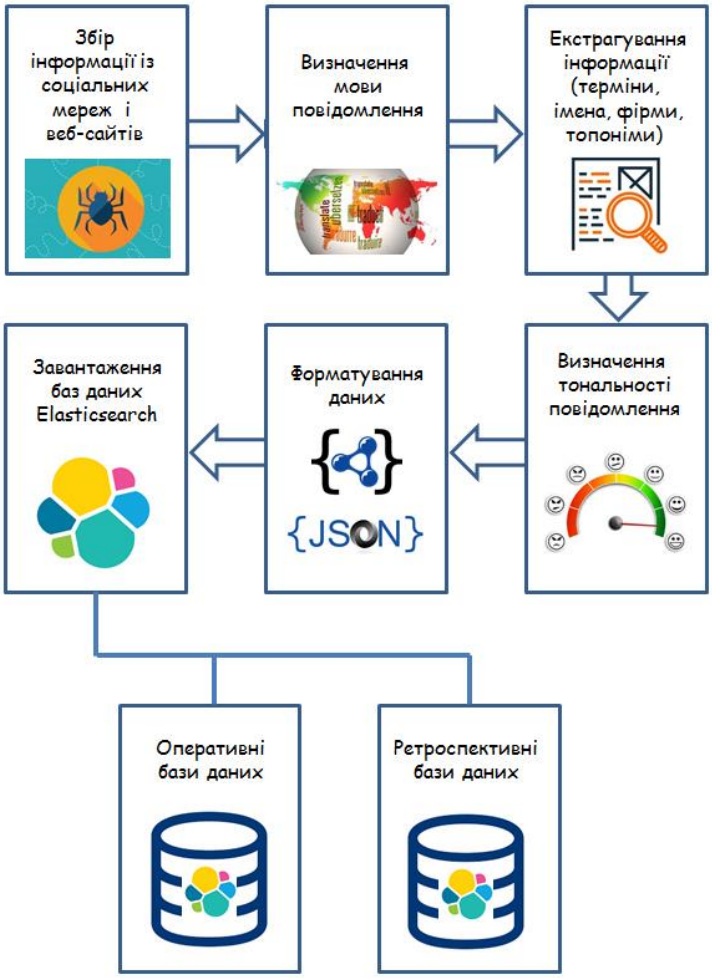

Рисунок 2: Етапи обробки інформації в системі «КіберАгрегатор»

Схема взаємозв'язку компонентів системи «КіберАгрегатор» складається з трьох основних частин - системного програмного забезпечення, системного ядра та програм користувача.

Інфраструктура системи включає:

- апаратне забезпечення (сервери, телекомунікаційне обладнання);

- операційна система (Linux);
- мови програмування та відповідні бібліотеки (Shell, JavaScript, Python, Perl, PHP);
- веб-сервер (Apache, nginx).
- ядро системи включає інструменти:
    - збір даних із соціальних мереж;
    - створення та ведення баз даних;
    - повнотекстовий пошук (система Elasticsearch, доповнена спеціальними засобами перетворення даних у форматі RSS);
    - аналітика та прогноз на основі вивчення мереж, статистики / динаміки тематичних інформаційних потоків (Kibana, Gephi, Matlab).

Крім того, передбачені програми користувачів:

- Веб-браузери;
- RSS-агрегатори (наприклад, FeedDemon 3.5, Feedreader 3.14, RSS Guard 3.4.1), які забезпечують доступ до баз даних «КіберАгрегатор» та опцій персоналізації (ведення персональних баз даних).

Особливості моделі, що розглядається, є одночасне використання методів та інструментів пошуку інформації, аналізу даних та агрегування потоків інформації.

Інтерфейс користувача

Система «КіберАгрегатор» забезпечує користувачеві веб-інтерфейс, з якого йому доступні функції пошуку та аналізу інформації (рис. 3).

Користувач системи отримує документи за запитом як у ретроспективній базі даних (Пошук), так і в поточній інформації (Поточна), а також для аналізу даних (Аналіз).

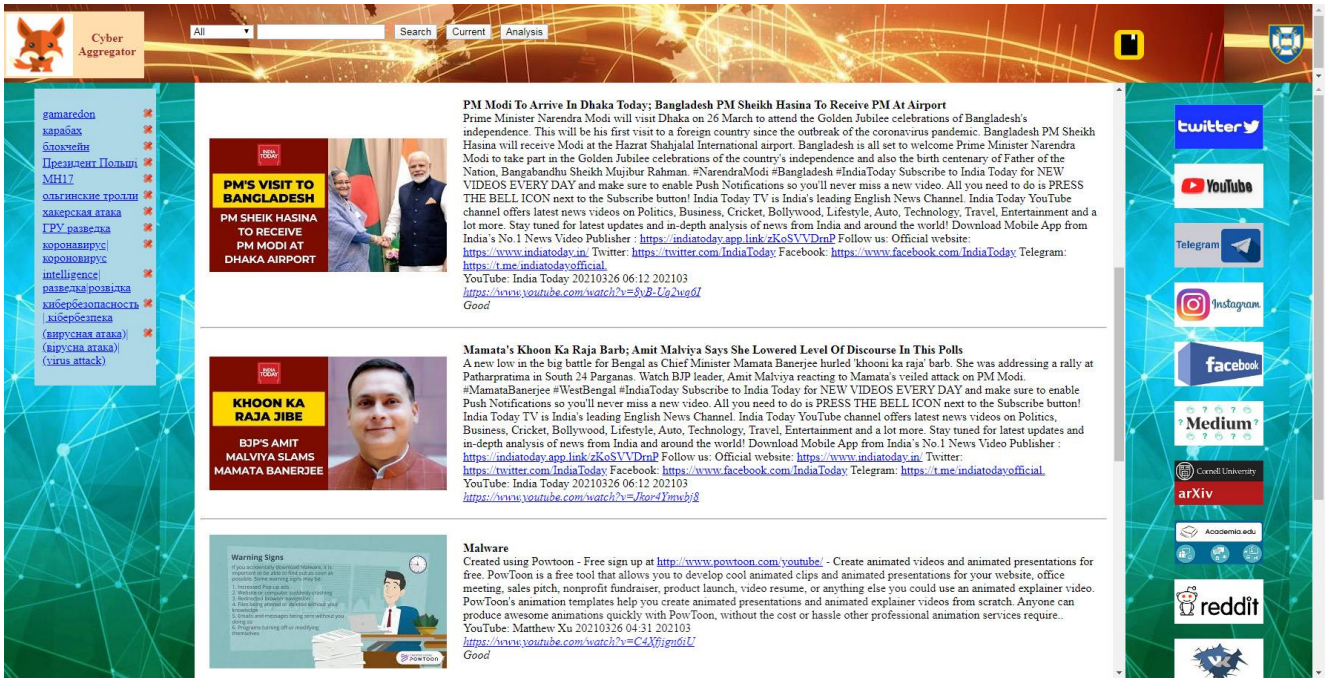

Рисунок 3: Інтерфейс системи «КіберАгрегатор»

Основним елементом інтерфейсу є дайджест найбільш актуальних повідомлень. В окремому блоці (Запити) відображаються збережені запити користувачів. Статистична інформація щодо заповнення бази даних системи з окремих соціальних мереж доступна у спеціальному розділі (Статистика джерел).

В результаті пошуку за запитом (рис. 4) користувачеві надається список відповідних заголовків повідомлень з гіперпосиланнями на повні тексти цих повідомлень у системі, а також на ці повідомлення в соціальних мережах.

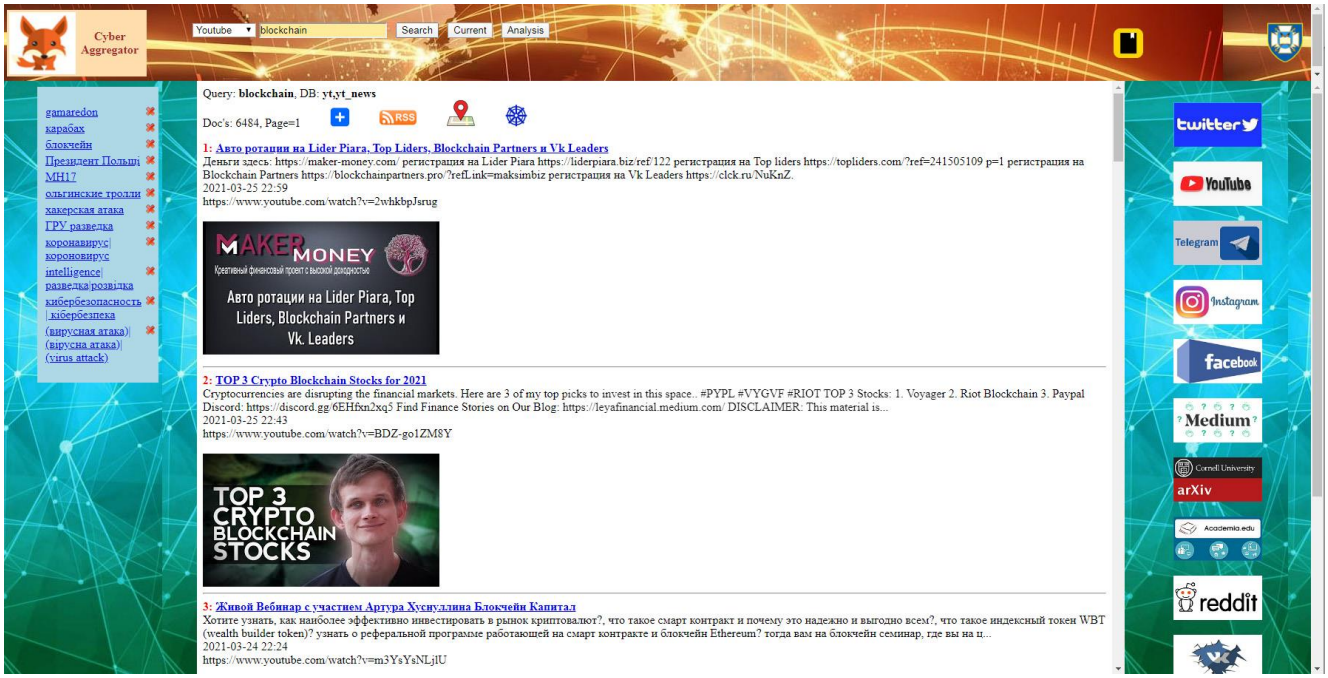

Рисунок 4: Фрагмент інтерфейсу користувача в режимі пошуку (результати пошуку за запитом «Блокчейн»)

Якщо запит створює документи, що відповідають інформаційним потребам, його можна зберегти для подальшого використання (Додати запит). Можна надалі виводити знайдені повідомлення у форматі RSS (з подальшим завантаженням цих результатів у так звані агрегатори RSS на постійній основі), а також відображати результати пошуку з деталями на географічній карті, яка масштабується як в автоматичному режимі, так і через налаштування (рис. 5).

В аналітичному режимі (Analysis) користувач отримує ряд інструментів, перший з яких являє собою графік (Graph), що відповідає часовому ряду кількості відповідних запитів повідомлень на день.

Користувачеві також надається можливість переглянути основні сюжети (Дайджест) по темі, кластери, згруповані за попередньо визначеними ключовими словами.

Система забезпечує режими формування мереж понять, що відповідають окремим повідомленням (людям, брендам), джерелам інформації (рис. 6). Ці режими дозволяють оцінити концепцію, дослідити взаємозв'язки між ними.

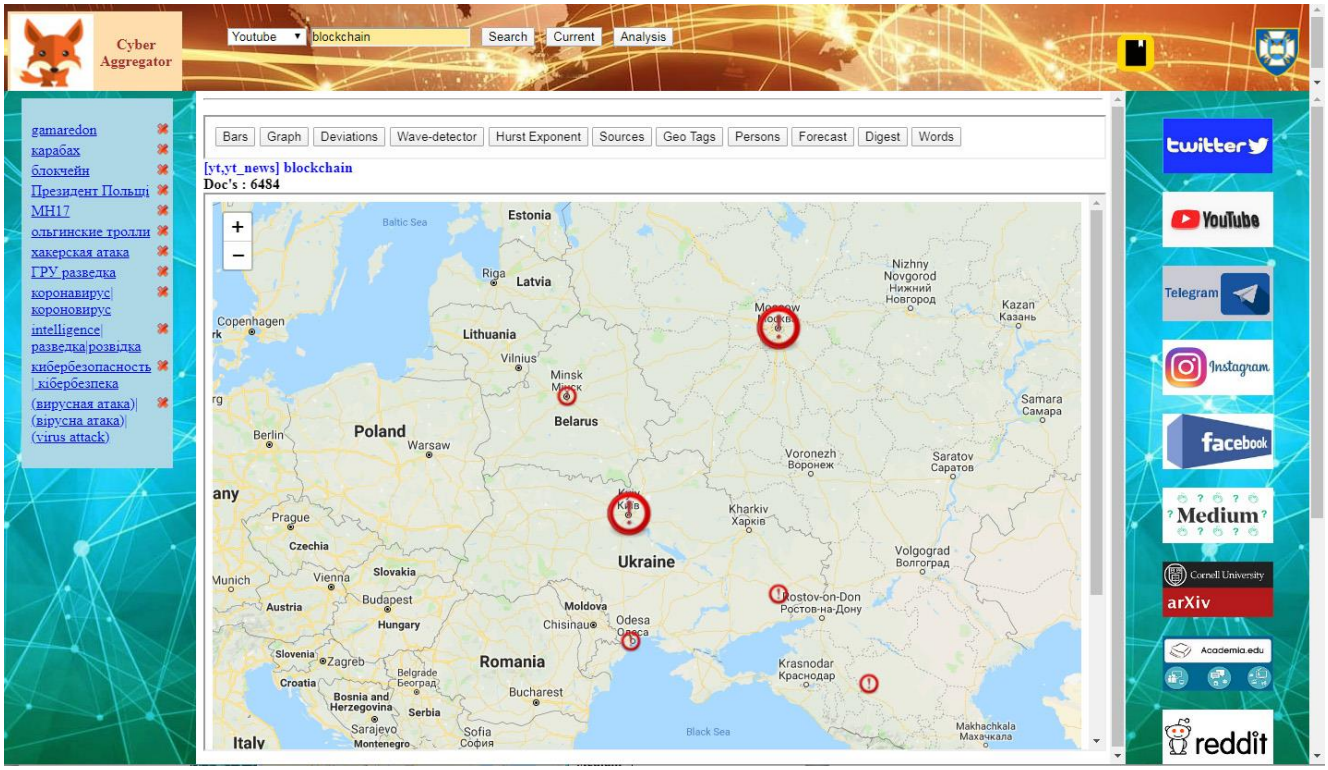

Рисунок 5: Фрагмент інтерфейсу геоінформаційної системи

Рисунок 6: Мережа взаємопов’язаних джерел інформації

В рамках макету реалізовано підхід до візуалізації тематичних кластерів, при цьому ставиться завдання візуалізації мережі термінів по відгуку пошукової системи в режимі реального часу – класифікатора, побудованого по мережевому принципу.

Модель динамічної класифікації інформації, яку можна розглядати як якусь «гру в слова» [16]. Саме ігровий принцип дозволив змоделювати навігатор, який в результаті знайшов своє застосування в реальному інтерфейсі, реалізованому на основі використання Javascript-бібліотеки D3.js [17]. Правила побудови мережі термінів як гри, яка відбувається на площині, розміченій шестикутними сотами, прості:

1 крок: в центральну соту вписується термін, відповідний деякому поняттю, яке найчастіше зустрічається у відгуку системи, що відповідає первинному запиту (Рис. 7).

2 крок: на підставі аналізу релевантного запиту масиву повідомлень із соціальних мереж вибираються 6 найбільш пов'язаних з першим терміном значущих термінів. Ці терміни вписуються у сусідні соти.

3 і наступні кроки: в вільні соти навколо кожної з заповнених сот вписуються найбільш пов'язані з заповненими сотами терміни (до 6, отриманих з того ж масиву). При цьому, якщо терміни вже були використані, то сусідні соти залишаються порожніми.

Процес зупиняється, коли додавання нових термінів стає неможливим.

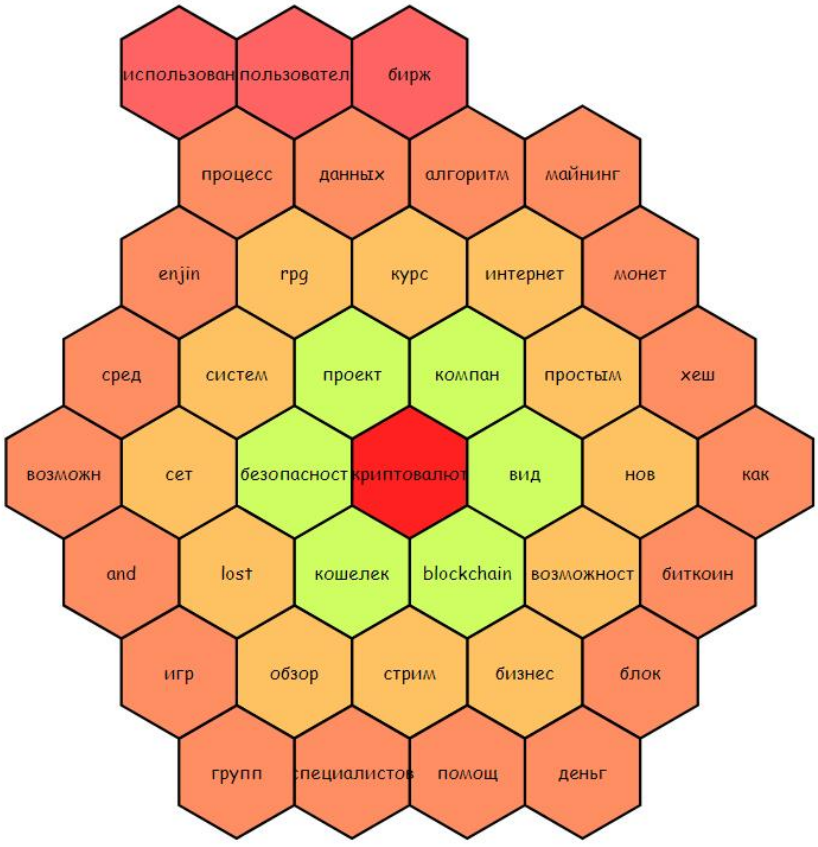

Рис. 7. Фрагмент первинної мережі термінів з масиву, відповідного слову із запиту

У наведеній на Рис. 1 мережі кожна клітинка-сота виступає як гіперпосилання. Активізація цього гіперпосилання викликає уточнення первинного запиту і призводить до виведення масиву релевантних документів.

Ілюстрація гри сотами цілком виправдана з двох причин: з одного боку, шестикутники щільно покривають площину, а з іншого, кількість вкладень в класифікаторі, що не перевищує 6, відповідає принципам ергономіки.

Зрозуміло, що для інформаційного наповнення моделі гри необхідний досить потужний інформаційний ресурс, який створено на бази системи «КіберАгрегатор» [4].

Режим "Аналітика" забезпечує можливість прогнозування (Forecast) за методом, запропонованим Д. Сорнетте, який базується на аналізі регулярності ринкових цін на товарних та фондових ринках до кризи. У роботах зазначається, що до кризи значення часового ряду (ціна) характеризується зростанням степеневого закону, ускладненим періодичними коливаннями, що сходяться до критичної точки, де ймовірність колапсу досягає максимального значення. Відповідна модель прогнозу, яка враховує лінійні часові періодичні коливання, має такий вигляд:

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

У цій моделі t_c – критичний час (кризовий час). Коефіцієнти моделі A, B, ω, φ визначаються за допомогою процедури підбору. Використовуючи модель Сорнета (клавіша прогнозу, рис. 7), можна отримати значення прогнозу для кількості відповідних мережевих публікацій на основі даних моніторингу.

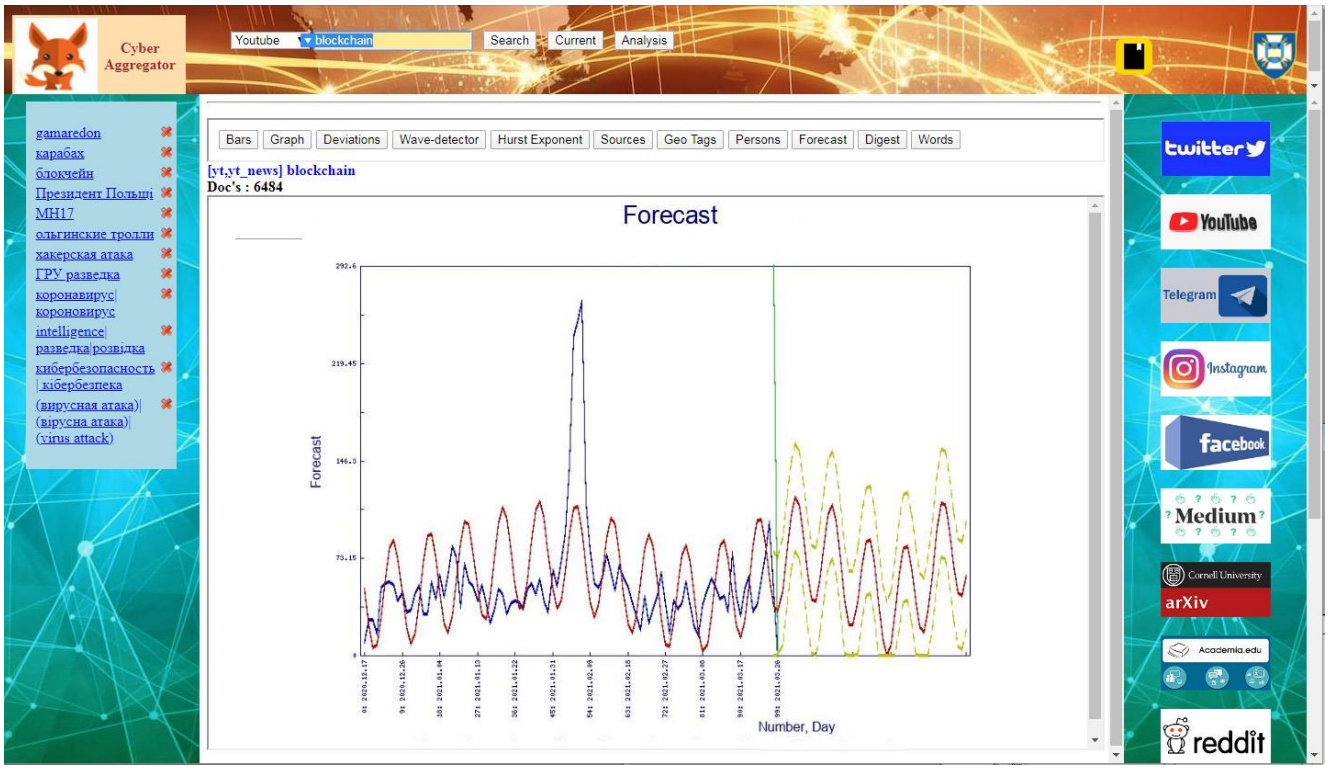

Рисунок 7: Рядок прогнозу за алгоритмом Сорне для часового ряду за запитом «Блокчейн».

Висновки

У цій роботі представлено основні розділи навчального курсу «Оброблення надвеликих масивів даних», основою якого є вивчення теоретичних засад і технологій Big Data. Основні всі розділи курсу, що розглядається, підтверджуються впровадженням в макеті, практичними роботами, які входять до складу курсу, що розглядався. Макет, який оформлено як систему «КіберАгрегатор», створено із застосуванням вільного системного програмного забезпечення вільно доступна.

Запропоновано та обгрунтовано інформаційні технології створення системи контент-моніторингу соціальних мереж з певних питань, відбору релевантної інформації із соціальних мереж, впровадження пошукової системи для їх доопрацювання користувачами, збереження запитів, проведенні аналітичних досліджень, прогнозування.

Звичайно, застосування технологій Big Data можливо для різних типів даних, не тількі для OSINT, на базі якої вирішується головна задача навчання в рамках цього курсу. Представлений макет у подальшому може бути розгорнутий для застосування в службовій інформаційно-аналітичній роботі.

Література

1. Укрощение больших данных / Бил Фрэнкс. – “Манн, Иванов и Фербер”, 2014. – 352 с.
2. Dmytro Lande, Ellina Shnurko-Tabakova. OSINT as a part of cyber defense system. Theoretical and Applied Cybersecurity, 2019. – N. 1. – pp. 103-108.
3. Основы Data Science и Big Data. Python и наука о данных / Дэви Силен, Арно Мейсман. – Питер, 2017. – 336 с.
4. Додонов А.Г., Ландэ Д.В., Прищепа В.В., Путятин В.Г. Конкурентная разведка - К.: ТОВ "Інжиніринг", 2021. - 354 с. ISBN 978-966-2344-79-0
5. NoSQL: новая методология разработки нереляционных баз данных / Прамодкумар Дж. Садападж, Мартин Фаулер. – И.Д. Вильяме, 2013. – 192 с.
6. Hadoop: Подробное руководство / Том Уайт. – Питер, 2013. – 672 с.
7. Elasticsearch, Kibana, Logstash и поисковые системы нового поколения / Пранав Шукла, Шарат Кумар. - Питер, 2019. – 363 с.
8. Learning Kibana 5.0. Exploit the visualization capabilities of Kibana and build powerful interactive dashboards / Bahaaldine Azarmi. - Packt Publishing, 2017. – 275 p.
9. Dmytro Lande, Igor Subach, Alexander Puchkov. System of Analysis of Big Data from Social Media. Information & Security: An International Journal 47, No 1 (2020): 44-61. DOI: doi.org/10.11610/isij.4703
10. Dmytro Lande, Oleksandr Puchkov, Ihor Subach. Система аналізу великих обсягів даних з питань кібербезпеки із соціальних медіа // Information Technology and Security. Том 8, N 1 (2020). - С. 4-18. DOI: doi.org/10.20535/2411-1031.2020.8.1.217993
11. Mastering Gephi Network Visualization. Produce advanced network graphs in Gephi and gain valuable insights into your network datasets / Ken Cherven. - Packt Publishing, 2015. – 378 p.
12. Network Graph Analysis and Visualization with Gephi Visualize and analyze your data swiftly using dynamic network graphs built with Gephi / Ken Cherven. - Packt Publishing, 2015. – 116 p.
13. Learning Neo4j. Run blazingly fast queries on complex graph datasets with the power of the Neo4j graph database / Rik Van Bruggen. - Packt Publishing, 2014. - 222 p.
14. Dmytro Lande, Igor Subach, Alexander Puchkov. System of Analysis of Big Data from Social Media. Information & Security: An International Journal 47, No 1 (2020): 44-61. DOI: doi.org/10.11610/isij.4703
15. Dmytro Lande, Oleksandr Puchkov, Ihor Subach. Система аналізу великих обсягів даних з питань кібербезпеки із соціальних медіа // Information Technology and Security. Том 8, N 1 (2020). - С. 4-18. DOI: doi.org/10.20535/2411-1031.2020.8.1.217993
16. Ландэ Д.В., Григорьев А.Н. Многоуровневый классификатор-навигатор по откликам информационно-поисковой системы // Компьютерная лингвистика и интеллектуальные технологии: труды международной конференции Диалог'2006 - Москва, Наука,2006. - С. 329-331.
17. Interactive Data Visualization for the Web. An Introduction to Designing with D3 / Scott Murray. – Published by O’Reilly Media, Inc., 2017. – 472 p.